UNESCO Conference on Education Data and Statistics
Session 1
7-9 February 2024

1 UIS/ESC/10

MEASURING AND MONITORING LEARNING OUTCOMES AND SKILLS: WHAT ARE THE CHALLENGES GOING FORWARD?

DRAFT

OCTOBER 2023

1. INTRODUCTION

This paper provides a comprehensive assessment and analysis of the status and gaps in measuring and monitoring learning outcomes and skills related to Sustainable Development Goal 4 (SDG4) - Quality Education. SDG4 aims to "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (UN, 2015, 2023).

Given the focus of this report on learning outcomes and skills, relevant frameworks, methodologies, and indicators related to SDG4 targets 4.1, 4.4, 4.6 and 4.7 and associated indicators will be explored in more depth (Table 1). This approach does not negate the underlying nature of other targets and indicators in relation to the main areas of concern of this paper. However, their exploration would require separate inquiry.

Measuring learning outcomes and skills is not trivial. The measurement of learning outcomes and skills can serve different stakeholders, e.g., parents, teachers, school leaders, inspectorate bodies, researchers, and policymakers. These stakeholders might seek information from measures of learning outcomes and skills for various purposes, including school choice, planning delivery, accreditation of studies, theory generation, system budgeting and reform, and other decisions such as rewarding, sanctioning, and training personnel (references). Investigating learning outcomes and skills can help, for instance, to decide where is best to spend the education budget, what teaching strategies support student learning in a given context (Clarke and Luna-Bazaldua, 2021), and whether and how schooling promotes equitable quality education leading to lifelong learning opportunities (reference). Depending on the goal and scope of measurement, specific data collection and data analysis methods can be employed. It is critical to remember that the Agenda 2030, from which the SDGs emanate, seeks "to fully engage in conducting regular and inclusive reviews of progress at sub-national, national, regional and global levels" (UN, 2015, p. 33). This means that various approaches to measurement might be found worldwide; still, there seems to be an intention to generate data for comparative purposes and benchmarking.

In the context of an initiative of the size of the Sustainable Development Goals (SDGs) and particularly, SDG4 on Quality Education, gaining an insight into the areas where progress has been achieved and those where further attention and improvement are needed is a critical, *albeit* challenging task to accomplish. This is so because the discussion on how to monitor and measure learning outcomes and skills is ongoing in nature. Previous international initiatives such as Education for All and the Millenium Development Goals faced their own obstacles in providing evidence of improvement in their different targets, and notably, the areas less amenable to quantitative measurements, including qualifiers such as *basic, minimum, enhancing,* among others, were particularly contentious (Torres, 1999; Unterhalter, 2014). This report suggests the SDGs might be subjected to similar caveats to those of previous programmes, either in terminology or operationalisation and therefore, identifying whether and how measurement evidence is been generated to tackle SDG4's targets is relevant for the way forward.

Measuring progress in learning outcomes and skills has several decades of development and debate. A review of the literature on the determinants of primary education outcomes in developing countries (Boissiere, 2004) identified that traditionally, advancements in educational psychology and sociology have been employed to understand educational outcomes; however, economic approaches and sophisticated statistical models have taken over recently. Therefore, different – sometimes antagonist – methods, make up the landscape regarding learning outcomes and skills measurement. Economic approaches, including those pertaining to Education Production Function theories, dominate, emphasising the role of a range of input-related aspects, such as student prior educational achievement, parental education and income, among others. Furthermore, processes, including teaching quality, school leadership, and outcomes, generally presented in standardised test scores in a limited number of subjects, e.g., language and

mathematics have been part of these approaches (Hanushek et al., 2016; Hanushek & Rivkin, 2006; Hanushek & Woessmann, 2010; Scheerens, 1991, 1997, 2015; Scheerens et al., 2003).

Table 1 SDG 4 targets and indicators related to learning outcomes

Indicator		Domain	Required definitions
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	Reading and mathematics	Minimum proficiency level Procedural quality minimum
4.2.1	Proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex	Learning, socio-emotional health	What is developmentally on track
4.4.2	Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills	Digital literacy skills	Relevant skills for employment, decent jobs and entrepreneurship
4.6.1	Percentage of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex	Literacy and numeracy	Fixed level of functional numeracy and literacy
4.7.4	Percentage of students by age group (or education level) showing adequate understanding of issues relating to global citizenship and sustainability	Global citizenship and sustainability	The definition of adequate understanding and what constitutes global citizenship and sustainability
4.7.5	Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience	Environmental science and geoscience	The definition of proficiency

1.1. Purpose of the position document

The paper aims to explore the existing frameworks, methodologies, and indicators that have been used for assessing learning outcomes and skills while identifying the areas that require further attention and improvement to measure and monitor the targets set by SDG4. More specifically, the objective is to make a meaningful contribution to the ongoing discussions and initiatives in this field, ultimately aiming to establish an international community of practice that can collectively address the challenges ahead.

2. Assessment of current status of measurement of indicator 4.1.1

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.
- **Sample:** the sample needs to be representative of the relevant population.
- **Procedures:** procedures need to comply with minimum standards of quality.

2.1. CHALLENGES

There are a few critical issues regarding reporting of indicator 4.1.1:

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 majority (89%) of countries end their primary cycle in Grades 5, 6, or 7 so the issue might be minor.

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 assessment helps comparability across countries, at one point in time. If each assessment programme produces statistics which are comparable over time, then statistics will be comparable across time and countries.
- *National* assessment programs are not comparable to each other by design, but they can still provide relatively reliable trend data if the measurement is of good enough quality.

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.

Procedural quality

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 operationalization.

Assessment implementation faces many methodological decisions including test formats and sampling decisions. There is no need for identical procedures and format across assessments. However, there is a need for a minimum set of procedures (procedural alignment), so data integrity is protected, and results

are robust as well as reasonably comparable for any given country over time, but also across countries at any given point in time.

Financial costs of assessments for countries

Assessments are relatively costly compared to other data collection systems such as EMIS. 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[1].

Low coverage of cross-national assessments specially in low-income and lower-middle income countries

SDG indicator 4.1.1 is being reported using various cross-national studies that are international (PIRLS, TIMSS) or regional (PILNA, SEA-PLM, PASEC, LLECE, SACMEQ) (Table 2). These tools have not been designed for SDG reporting but, in 2018, the Global Alliance to Monitor Learning (GAML) and the Technical Cooperation Group on SDG 4 indicators (TCG) agreed that these assessments could be used to report learning based on their proficiency levels that "mapped" best to the global MPL.

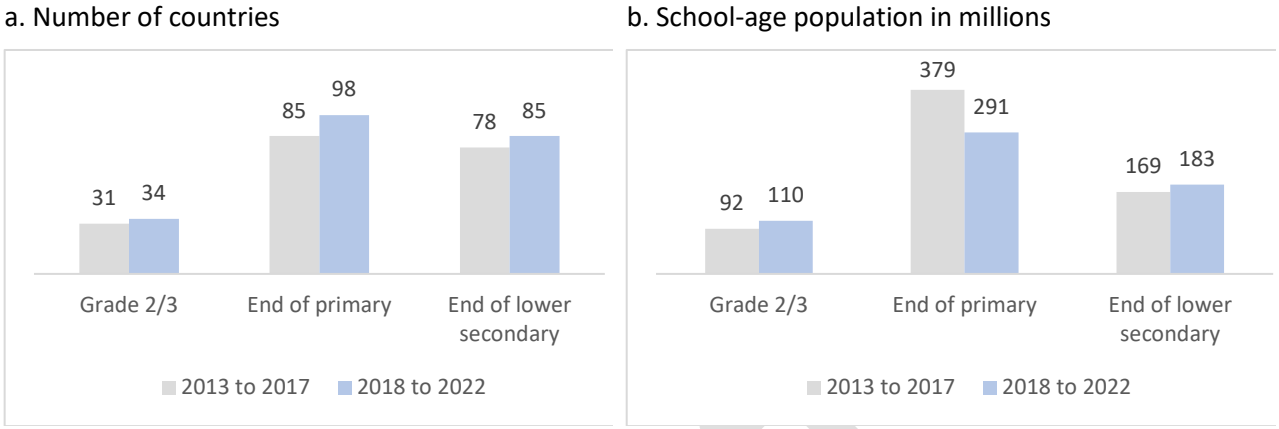
Table 2 Assessment programs by grade or age and use for reporting on SDG indicator 4.1.1

Grade	International assessment program
School-based	
SDG 4.1.1a: Early grades	
2	EGMA, EGRA, **PASEC**
3	EGMA, EGRA, **ERCE**, **AMPLa**
SDG 4.1.1b: End of primary	
4	**PILNA**, **LaNA**, **PIRLS**, **TIMSS**
5	**SEA-PLM**
6	**LaNA**, **PASEC**, **PILNA**, **SACMEQ**, **ERCE**, **AMPLb**
SDG 4.1.1c: End of lower secondary	
8	**TIMSS**
Age/ 15 years	**PISA**

However, the production of comparable learning outcomes is not progressing fast and equally enough. Regardless of the coverage criterion (number of countries or population), coverage is much higher at the end of primary and end of lower secondary than for grades 2 or 3 (Figure 1).

[1] For more information on costing, refer to this paper: https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2023/05/Countrys-reporting-option_Zambia_2023.05.15_FINAL.pdf

Figure 1 - Coverage of learning assessments, by level of education



National assessments: alignment constraint and uncertainty on procedural quality

While data from many national learning assessments are readily available, every country sets its own standards, leading to inconsistent definitions of performance levels. Analysis of results therefore remains contained to one test, methodology and scale.

While methodologies tend to converge between international and regional assessments, it is still difficult to situate assessments in a common reference level national assessment and there is uncertainty with respect to the set of procedures utilized for sampling, data management and reporting.

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 operationalization.

Assessment implementation faces many methodological decisions including test formats and sampling decisions. There is no need for identical procedures and format across assessments. However, there is a need for a minimum set of procedures (procedural alignment) so data integrity is protected, and results are robust as well as reasonably comparable for any given country over time, but also across countries at any given point in time.

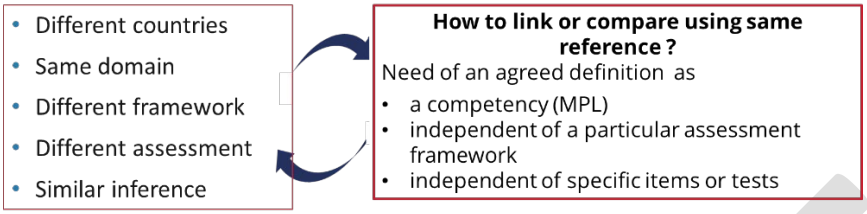
2.2. DEVELOPMENTS AND PENDING CHALLENGES

Achieving statistics that are comparable across programmes and countries is perhaps more difficult than is assumed. This is because different regions have different traditions concerning the stringency of proficiency benchmarks at different grades. Moreover, these realities further complicate comparisons across countries, which often involve comparing slightly different grades, even at the same educational level. The focus of UIS work has been the definition of the minimum proficiency level as the concept that allows the reporting, and a set of linking strategies to the proficiency framework.

The UIS has decided alignment to a concept instead of to a score in points to produce comparable data across programmes to the minimum proficiency level, strategy that requires an alignment strategy to express all assessments in that benchmark summarized in Figure 2.

The task implied the creation of a set of global standards in a time-consuming but necessary process. This would make it possible for disparate assessments to refer to standards without having to carry out the same assessment. The following suite of tools would make this possible:

Figure 2 – Linking or comparing using the same reference



2.3. STANDARDS

2.3.1. *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. The MPL is a reference point for reporting on minimum competencies at each schooling level, without requiring a single test to solve comparability.

The Proficiency Level Descriptor (PLD) of the MPL is the key standards for each grade and domain that allows the use of multiple assessment to report for the indicator. Table 3 presents the global MPL definitions for the domain of mathematics and reading.[2]

The first step in the implementation of the MPL was to agree with the cross-national assessment programs on the PLD in each assessment program better aligned to the MPL PLD. This step agreed in 2018 needs further steps in validation through a standard setting exercise as the assessment programs have not been designed to measure SDG4 and then they are used by:

- approximate the grade levels of interest for reporting;
- use assessment program existing proficiency levels;
- identify the PLD that is better aligned to the minimum proficiency levels;
- **use that level to report** until the standard setting exercise is finalized**;**
- build national technical capacity but do not directly address national assessment development.

Table 3. Minimum proficiency levels for reading and mathematics - Indicator 4.1.1

Educational Level	**Descriptor**	
	Reading	**Mathematics**
Grade 2	They read and comprehend most of written words, particularly familiar ones, and extract explicit information from sentences.	Students demonstrate skills in number sense and computation, shape recognition and spatial orientation.
Grade 3	Students read aloud written words accurately and fluently. They understand the overall meaning of sentences and short texts.	

[2] 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.

Grade 4-6	Students interpret and give some explanations about the main and secondary ideas in different types of texts. They establish connections between main ideas on a text and their personal experiences as well as general knowledge.	Students demonstrate skills in number sense and computation, basic measurement, reading, interpreting, and constructing graphs, spatial orientation, and number patterns.
Grade 8 & 9	Students establish connections between main ideas on different text types and the author's intentions. They reflect and draw conclusions based on the text.	Students demonstrate skills in computation, application problems, matching tables and graphs, and making use of algebraic representations.

Source: UNESCO Institute for Statistics, 2021, p. 4

2.3.2. THE GLOBAL PROFICIENCY FRAMEWORK (GPF)

The Global Proficiency Framework is a useful global reference defining proficiency levels for reading and mathematics that learners are expected to demonstrate at the end of each grade level, from grades 1 to 9, as common reference. The four levels outlined in the GPF – Below Partially Meets, Partially Meets, Meets, and Exceeds Global Minimum Proficiency – form a common scale from low to high achievement.

An additional standard has been created. The Global Proficiency framework and its related proficiency levels give guidance as to the minimum set of skills that students should acquire on the pathway to mastery of reading and mathematics.

2.3.3. A SET OF TOOLS TO LINK ASSESSMENT PROGRAMS TO THE MPL

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 assessments programs built for completely different purposes to express in a way that allow 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.

Different solutions have been suggested to obtain data that can be used to measure and monitor SDG 4.1 and its indicators. Three of the most widely discussed and supported by the international assessment community are the Rosetta Stone project, the Policy Linking Methodology, and the Assessments for Minimum Proficiency Levels (AMPLs). In what follows, we provide some basic information about each of these initiatives, as well as their main strengths and limitations. This information will feed into the concluding section of this document that will be a proposal of an agenda for the international educational measurement community to follow in the short and mid-terms. Please refer to the paper on country’s options to report on 4.1.1 for more information on these different linking strategies, and a summary of costs, benefits, status of execution, milestones executed and pending, and time frame.

Rosetta Stone

One of the initiatives that has been tested to try to harmonize data from different assessments is the **Rosetta Stone** project that was led by the International Association for the Evaluation of Educational Achievement (IEA). The Rosetta Stone Study is designed to measure global progress towards SDG 4.1.1 by relating national and regional learning assessments to international learning assessments. It is named after the famous archaeological discovery that enabled translation between different written languages: the Rosetta Stone. The goal of the study is to provide countries that participated in regional or national assessments but not in international assessments 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.

In a first effort to implement this approach and establish concordance tables, the regional assessments Regional Comparative and Explanatory Study (ERCE) and Programme d 'analyse des systèmes éducatifs de la CONFEMEN (PASEC) are linked to two international assessments of IEA, namely Trends in International Mathematics and Science Study (TIMSS) for mathematics and Progress in International Reading Literacy Study (PIRLS) for reading.

Policy Linking

Another approach related to the harmonisation of assessments is the **Policy Linking** methodology (UNESCO Institute for Statistics, 2021) which is a non-statistical method that uses judgment to align and match items from the national assessment with the Global Proficiency Framework (GPF). This process establishes the internationally comparable global benchmarks based on the descriptors of each benchmark specified in the GPF.

Three major tasks – alignment, matching and setting benchmarks – have to be done in a workshop of 5-6 days with 15-20 panelists (teachers), curriculum and assessment experts of each grade/subject, to identify and set, if feasible, the required benchmarks for international reporting on SDG indicator 4.1.1.

To produce reliable benchmarks for international reporting, the Policy Linking Toolkit (PLT) specifies five criteria including: a sufficient number of national items are aligned with the GPF; samples are nationally representative; and national assessments are administered according to minimum quality standards. If criteria are not met, the workshop will be considered a capacity building activity.

The Policy Linking methodology was proposed during the Global Alliance to Monitor Learning (GAML) meeting in 2017, piloted in 2019 for the first time, revised in 2020 and piloted again in 2021–2022. The PLT was then revised in January 2023, and it is under piloting phase.

Assessments for Minimum Proficiency Levels (AMPLs)

A third method that has been proposed as a solution is the Assessments for Minimum Proficiency Levels (AMPLs). AMPLs are robust tools targeted at measuring the attainment of a single proficiency level for each of the reading and mathematics domains at a given level of the education cycle. AMPL tools allow to identify the proportion of children and young learners in each level of education who are achieving at least the Minimum Proficiency Level (MPL). This allows countries the production of international comparable learning outcomes data to report on the global indicator SDG 4.1.1.

AMPL-b is the first AMPL developed in 2021 in both English and French and was implemented in six African countries as part of the MILO project in 2021 - Burkina Faso, Burundi, Côte d'Ivoire, Kenya, Senegal and Zambia. AMPL-b was administered as a standalone module in Sierra Leone in 2022 and it is scheduled to be implemented in Jordan and Pakistan.

AMPL-a, which measures proficiency in early grades, is under development and will be piloted and administered in 2023 in both languages English and French.

2.3.4. COUNTRIES: ALTERNATIVES FOR REPORTING.

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 known as the Learning Data Compact.

A set of 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. Those principals are to build on what exists; allow flexibility to ensure alignment with country needs (not one-size-fits-all); foster country ownership through a demand-driven approach; ensure data is relevant for decision-making.

A national assessment 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. To ensure that assessments can accurately monitor progress for decision making, data also must be internationally comparable. 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.

Countries' options to report are reflected in table 4, but the choice should be guided by what assessment(s) are fit-for-purpose and most cost-effective for them, taking into account country's initial situation and the objective to have comparability over time and representativeness of results at the national level.

Two special cases should be noted. National assessments could be used to report subject to the use of statistical linking that could be implemented using calibrated modules such as AMPL. Other tools such as the Minimum Proficiency Levels and the Global Proficiency Framework serve to understand and benchmark to global standards, while Policy Linking (PL) serves to engage national stakeholders and analyse the assessment vis-à-vis those standards. The UIS considers countries would like to report globally on indicator 4.1.1: they add a calibrated module to the national assessment, such as AMPL and supplement this with Policy Linking (described above) for capacity building purposes, given the methodology is still under piloting phase.

A second case is related to early grades, or 4.1.1a, where the existence of tools take particular relevance - such as the Early Grade Reading/Mathematics Assessment (EGRA/EGMA), the PAL Network citizen-led assessments, and UNICEF's Foundational Learning Module of its MICS household survey - and that potentially could serve to report. Even though these assessments have been applied globally, they cannot be currently used for global reporting, mostly because they were not intended to generate comparable data. Nevertheless, they do have the potential to be used for global reporting and the UIS is looking into how to make the best use of such assessments. For more details, please refer to this blog here.

Table 4 Alternatives for country reporting SDG indicator 4.1.1

	4.1.1.a	4.1.1.b	4.1.1.c	Coverage
National assessments - statistical linking through calibrated modules				
AMPL	•	•		
PISA module			•	
Participate in a Cross-National assessment				
PILNA		•		Pacific islands
PASEC	•	•	•	Mainly Africa (Francophone)
SACMEQ		•		Africa (Southern and Eastern)
SEA-PLM		•		Southeast Asia
LLECE	•	•		Latin America
TIMSS	•	•	•	Global
PIRLS	•	•		Global
PISA			•	Global

3. Assessment of current status of measuring indicator 4.4.2, 4.6.1, 4.7.4 and 4.7.5

Significant progress has been made in the establishment of methodological frameworks for these indicators. The development of these frameworks provides a structured approach, offering guidelines and principles for data collection and analysis. However, despite the defined methodological frameworks, a

substantial problem exists in the form of low data coverage, particularly in Low- and Middle-Income Countries (LMICs). Addressing this challenge requires a concerted effort to establish common definitions and metrics, ensuring a standardized approach across assessments (Table 5).

Table 5 Main parameters of assessment availability

Indicators	Methodological framework	Data sources	Admin. last	Cycle length	Coverage countries	Coverage population (%)
4.4.2	Yes	Skills' assessment surveys of the adult population (PIAAC)	2017	undefined	5	2
4. 6.1	Yes				7	3
4.7.4	Yes	ICCS	2016	6/7 years	23	10
4.7.5	Yes	TIMSS, PISA	2019/2022	4/5 years	38	16

3.1. Pending Issues

The fact is that LMICs often lack the resources and infrastructure needed to develop and implement tools to measure these indicators effectively. Without enough data on these indicators, it becomes difficult for these countries to identify specific areas for improvement and allocate resources efficiently. Moreover, this data gap inhibits international efforts to provide targeted support to the countries that need it most, hindering the global progress towards achieving SDG 4. Addressing this issue necessitates not only the development of appropriate measurement tools but also targeted capacity-building initiatives in LMICs to ensure that these indicators are comprehensively and accurately measured, providing a foundation for informed decision-making and policy formulation in the realm of education.

4. Methodology for Harmonizing Context Questionnaires

Beyond harmonising data on learning outcomes and skills, harmonising context questionnaires from different large-scale assessments is critically important for enabling robust comparative analyses of trends, patterns, and determinants of educational inequality across countries and over time. Such data harmonisation would allow researchers to assemble large longitudinal datasets that could provide novel insights into key issues like learning inequality, school segregation, privatisation, and performance in LMICs. By collaborating across institutions and drawing on existing assessments from various contexts, harmonised questionnaires could unlock opportunities for impactful research that is highly policy-relevant and contributes directly to monitoring progress on different indicators of SDG 4.

While the primary objective of CNAs is to estimate measures of learning outcomes for a country, they also collect a rich set of background information about teachers, students and schools. From students, CNAs typically collect information about their experience at schools, their attitudes towards subjects being taught, and the characteristics of their parents and households in addition to core demographic information of age and sex (Table 6). From teachers, CNAs collect information about their attitudes towards teaching, their opinions about teaching resources, their educational background and on-going professional development, and from schools, CNAs collect information about infrastructure, location and opinions from the school directors about the availability of resources at school and how they interact with parents. For schools, there is some variation in how objective the data collected is.

Table 6. Typical questionnaires and data collected in CNAs

Cognitive test	Test items (questions) for measuring learning outcomes
Student questionnaire	Basic demographic information (sex, age) Household and socio-economic background School-related experiences (including exposure to bullying) Learning-related experiences (classroom activities) Self-perceptions, interests and aspirations related to different subjects Use and proficiency of ICT
Teacher questionnaire	Demographic and background information (sex, age, years teaching, subjects taught) Qualifications and training Types of teaching practices used and challenges faced
School director questionnaire	Demographic and background information (sex, age, years of experience) Qualifications and education School characteristics Opinions about availability and adequacy of resources Management and governance Interaction with parents and school communities Challenges faced in teaching

It is from these questionnaires that it can be determined whether an SDG indicator can be estimated or not. It is also these questionnaires that determine what sub-populations the indicators can be estimated for and how they inform the equity dimension. Generally, these include urban and rural location of the school, socio-economic status (SES) of the student (relative to other students, not the population), and sex of student (or teacher)[3].

4.1. CONTEXT QUESTIONNAIRES - HARMONIZATION

Interpreting and deriving policy advice poses significant challenges due to the inherent differences in definitions across various dimensions, such as rural/urban classifications, socio-economic status (or wealth), period of reference, and more. The disparities in these definitions hinder the comparability of assessment outcomes and data interpretation: for instance, what constitutes 'rural' in one country might differ from another. Similarly, varying definitions of SES can impact the analysis of disparities in educational outcomes among different social or economic strata. Bridging these definitional gaps requires international collaboration and the development of standardized frameworks that ensure uniformity in definitions, enabling accurate assessments and facilitating meaningful comparisons across regions and socioeconomic contexts.

The harmonisation of context questionnaires can also serve other various valuable purposes, including:

- *Comparative Research:* Harmonised questionnaires would allow researchers to compare educational contexts across countries, regions, or over time. This can help identify trends, similarities, and differences in educational systems, policies, and practices.
- *Policy Analysis:* Policymakers could use harmonised data to evaluate the effectiveness of educational policies and interventions by comparing outcomes across different contexts. This aids in evidence-based policymaking.

[3] For more information, please refer to this paper by the UIS: 'Monitoring of the Sustainable Development Goals using Large-Scale International Assessments' (2022).

- *Equity Analysis:* Researchers could use harmonised data to investigate educational inequalities, including disparities in access to resources and opportunities. This could inform efforts to reduce educational inequities.
- *Curriculum Development:* Harmonised context data can help in the development of curricula that are better aligned with the needs and challenges of students across different contexts.
- *Teacher Training:* Understanding the contextual factors affecting teaching and learning can inform teacher training programs, ensuring that educators are well-prepared for the specific challenges they may face.
- *Resource Allocation:* Governments can use harmonised data to allocate educational resources more effectively, targeting areas with the greatest need.
- *Cross-Cultural Research:* Researchers can conduct cross-cultural studies to explore how cultural factors impact education and learning outcomes.

4.2. Challenges in harmonization

The challenges are briefly described and all demand adaptation to be relevant and context sensitive:

1. *Cultural and Linguistic Differences*: Variations in cultural and linguistic factors can lead to diverse interpretations of existing questions and responses needing cultural adaptation.
2. *Contextual Specificity*: Educational contexts vary widely between countries and regions.
3. *Differing Educational Systems*: Differences in educational systems and policies between countries demand the harmonisation of questions that were not originally designed to be universally applicable.
4. *Data Collection Methods*: Countries may employ different methods, procedures, and instruments for data collection. It is complex to harmonise these in existing questionnaires while maintaining data quality.
5. *Response Variability*: Individuals' responses to existing questions may vary based on cultural norms and expectations. Consistent interpretation and response patterns during the harmonisation process is relevant.
6. *Privacy and Ethical Considerations*: Existing data collection often involves sensitive information about students, teachers, and schools. Ensuring data privacy and adhering to ethical guidelines while harmonising can be a complex issue.
7. *Quality Control*: Maintaining data quality and consistency across diverse contexts when harmonising existing questionnaires requires rigorous quality control measures, which can be resource intensive.
8. *Changing Contexts*: Educational contexts evolve over time, and existing context questionnaires may become outdated demanding regular to keep the questionnaires relevant during harmonisation.
9. *Political and Cultural Sensitivities*: Some questions in existing questionnaires may touch on sensitive political or cultural issues calling for a common ground and agreement on appropriate wording.
10. *Data Standardisation*: Harmonising data formats and coding schemes can be a technical challenge then dealing with existing questionnaires, especially when countries use different systems.

5. Setting an Agenda for the Measurement Community

5.1. A Blueprint to guide country reporting and ensure quality and alignment – 4.1.1

The existing protocol for reporting and menu of options including the alternative ways of reporting and menu of options available to ensure users are fully aware of the properties and limitations of the data.

While initiatives like the Rosetta Stone, Policy Linking, and AMPLs have worked to harmonize different educational assessments, a standardized blueprint is still needed to systematically evaluate which assessments are suitable to include in these harmonization efforts. As more national and international

large-scale assessments emerge, having clear criteria to analyze their quality, comparability, and viability for harmonization is critical.

A comprehensive blueprint should outline key factors to examine for each assessment under consideration. For example:

- Alignment to learning standards and frameworks - Assessments must adequately measure the intended curriculum and skills.
- Psychometric properties - Evidence of reliability, validity, appropriate difficulty, discrimination, etc.
- Representativeness - Samples must reflect target populations.
- Comparability of administrations - Consistent, standardized administration procedures.
- Transparency of processes - Assessment design, sampling, analysis should be well documented.
- Capacity for linking - Enough equivalent items/proficiency levels to enable linking.
- Stakeholder involvement - Inclusion of experts throughout design and implementation.
- Feasibility of participation - Reasonable costs, schedules, and burdens for countries.

A detailed blueprint incorporating these elements will be developed on an annex of this position paper and will represent the first point for the agenda to be proposed to the international educational assessment community. This blueprint will allow for rigorous vetting of assessments to determine their appropriateness and technical capacity for harmonization initiatives. Global standards and participation can then be strengthened. For example, developing an agreed-upon model through the GAML network should be pursued.

5.2. HARMONIZATION OF CONTEXT QUESTIONNAIRES

Harmonising existing context questionnaires in international large-scale assessments requires careful consideration of these challenges to ensure that the resulting harmonised data is reliable and comparable across different educational contexts.

The main strategy would be the creation and systematic maintenance of a harmonised dataset of international large-scale assessments in education, that provides longitudinal indicators of levels and trends in educational achievement (see previous section) and its potential determinants, at the country level. This harmonised dataset would include indicators from national, global and regional large-scale school assessments that meet the minimum quality requirements established by a blueprint created for this purpose (see next section). The idea would be to ensure that only assessments that contain information on student background, school characteristics, and learning outcomes that is comparable over time as well as between countries are included in the harmonisation. Previous studies have developed harmonisation methodologies (Angrist et al., 2021; Gust et al., 2022; Patel & Sandefur, 2019; Sandoval-Hernandez, 2022). The harmonisation of context questionnaires should build on this work by creating indicators for student, teacher, and head-teacher background, school resources, and educational inequality.

5.3. PROPOSED SOLUTIONS TO ENHANCE COVERAGE - 4.6.1

- Consideration of alternatives to replace PIAAC indicator, such as literacy rate that has a high rate of coverage and frequency in reporting.
- Utilization of artificial intelligence for producing indicators for adult population (e.g., reading and scoring available bodies of text).