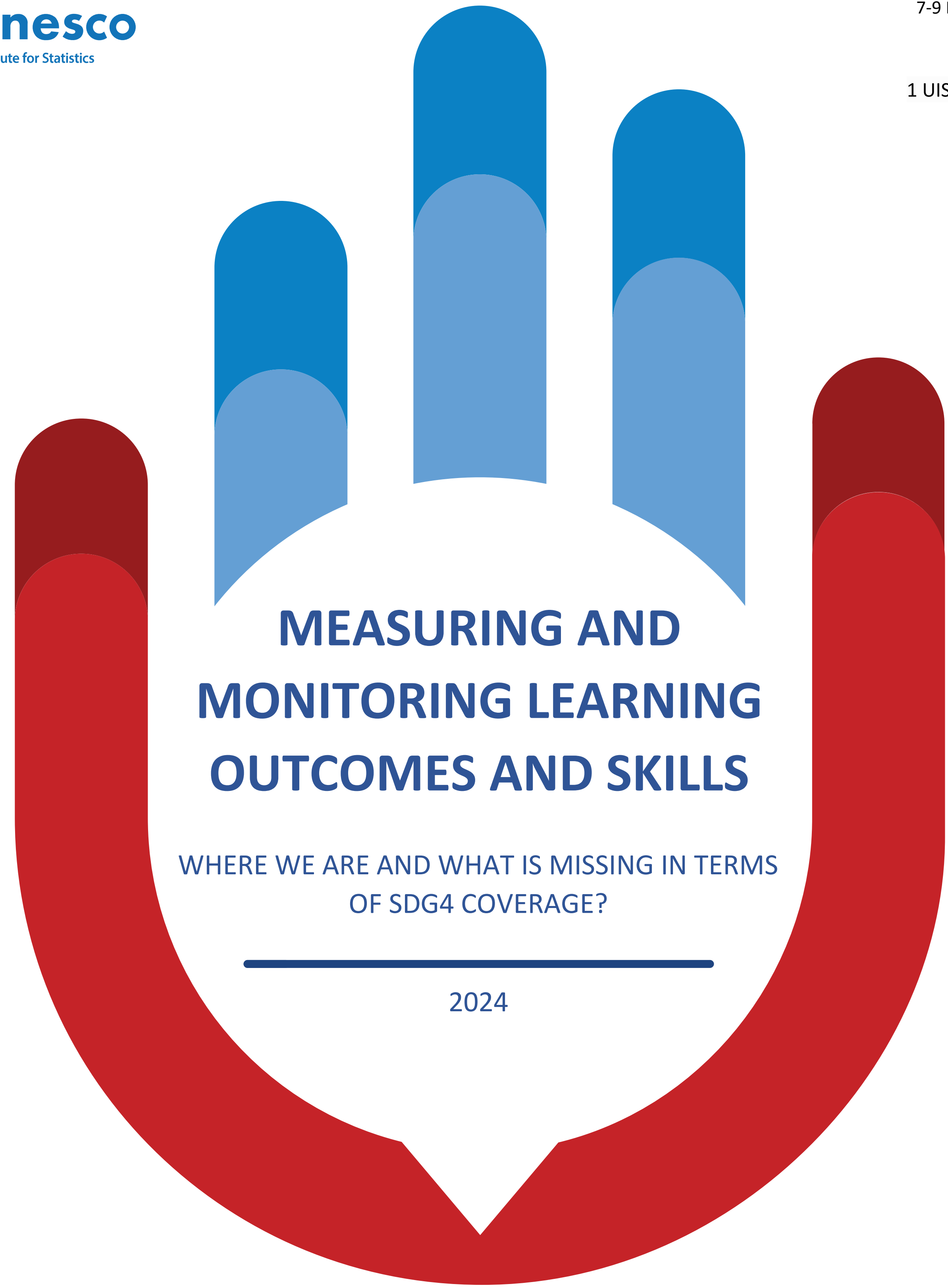
Institute for Statistics

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

1 UIS/EDS/BP.LO.1



# MEASURING AND MONITORING LEARNING OUTCOMES AND SKILLS

## WHERE WE ARE AND WHAT IS MISSING IN TERMS OF SDG4 COVERAGE?

2024

2024 CONFERENCE ON
EDUCATION DATA
AND STATISTICS

## Table of contents

1. Introduction .......... 3
Background and significance of measuring learning outcomes and skills .......... 4
Purpose of the position document .......... 5
2. Assessment of SDG4 Coverage .......... 5
Current status in measuring SDG4 indicators related to Learning Outcomes and Skills. .......... 5
Some of the challenges in measuring SDG4 indicators related to Learning Outcomes and Skills. .......... 7
3. Harmonisation of different assessments and data sources on educational achievement .......... 9
Rosetta Stone .......... 10
Policy Linking .......... 10
Assessments for Minimum Proficiency Levels (AMPLs) .......... 11
4. Harmonisation of context questionnaires from different assessments .......... 11
Overview of existing context questionnaires in different assessments and potential uses .......... 11
Challenges in harmonizing these questionnaires .......... 13
Proposed strategies to achieve harmonization .......... 14
The Need for a Blueprint to Evaluate Assessments for Harmonization .......... 14
References .......... 16

# 1. Introduction[1]

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). This goal is integrated by a series of targets and indicators as shown in Table 1.

*Table 1: SDG4 Targets and indicators*

| |
|---|
| **4.1** By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes |
| 4.1.1<br>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<br>4.1.2<br>Completion rate (primary education, lower secondary education, upper secondary education) |
| **4.2** By 2030, ensure that all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education |
| **4.3** By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university |
| **4.4** By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship |
| 4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill |
| **4.5** By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations |
| **4.6** By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy |
| **4.7** By 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development |
| 4.7.1 Extent to which (i) global citizenship education and (ii) education for sustainable development are mainstreamed in (a) national education policies; (b) curricula; (c) teacher education and (d) student assessment |

[1] Document prepared for the UNESCO Conference on Education Data and Statistics 2024, by Andrés Sandoval-Hernández (University of Bath) and Artemio Cortez-Ochoa (University of Bristol).

| |
|---|
| **4.A** Build and upgrade education facilities that are child, disability and gender sensitive and provide safe, nonviolent, inclusive and effective learning environments for all |
| **4.B** By 2020, substantially expand globally the number of scholarships available to developing countries, in particular least developed countries, small island developing States and African countries, for enrolment in higher education, including vocational training and information and communications technology, technical, engineering and scientific programmes, in developed countries and other developing countries |
| **4.C** By 2030, substantially increase the supply of qualified teachers, including through international cooperation for teacher training in developing countries, especially least developed countries and small island developing states |

*Source: (UN, 2015).*

Given the focus of this report on learning outcomes and skills, relevant frameworks, methodologies and indicators related to SDG4 targets 4.1, 4.4 and 4.7 and associated indicators will be explored in more depth (denoted in blue background in 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.

## Background and significance of measuring learning outcomes and skills

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).

### 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 in order 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 SDG4 Coverage

### Current status in measuring SDG4 indicators related to Learning Outcomes and Skills

Based on the Official List of SDG 4 Indicators (March 2022).

*SDG 4.1 and indicators*

| **4.1** By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes |
|---|
| 4.1.1<br>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<br>4.1.2<br>Completion rate (primary education, lower secondary education, upper secondary education) |

Currently, the overarching target 4.1 has been operationalised via indicator 4.1.0 Proportion of children/young people prepared for the future by sex (UNESCO, 2017). This indicator embraces completion (primary or lower secondary), and achievement of minimum proficiency (reading and mathematics), by sex when both completion rates and achievement data include such level of disaggregation (UNESCO Institute for Statistics, n.d.-b). Regarding completion data, the Adjusted Bayesian Completion Rates (ABC) Estimation (Dharamshi et al., 2021) has been used as it addresses the challenges known to administrative and household surveys and provides time series for 157 countries. Administrative surveys, for instance, might not be available in all contexts, while household surveys might provide conflicting information between waves and fail to account for those who complete school later due to repetition or start their education later (UNESCO Institute for Statistics, n.d.-b).

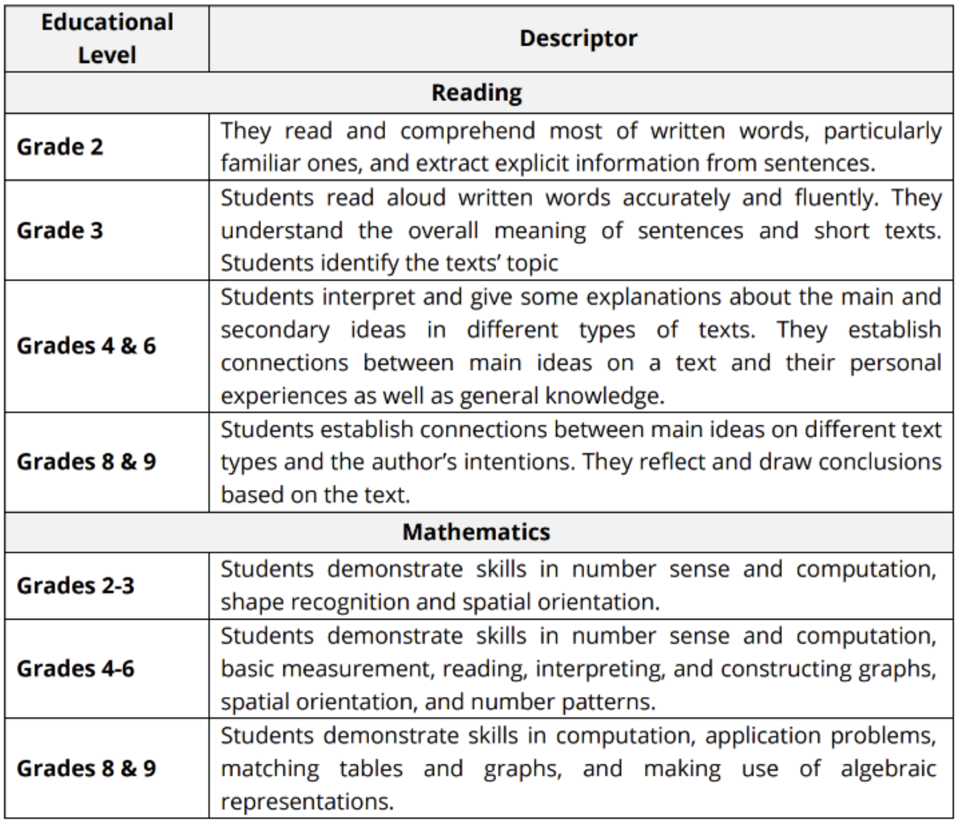
According to its creators, "The objective of the ABC model is to consolidate observations from different surveys, provide estimates for years without a survey, and allow for short-term 'now-casts' of current completion rates." (Dharamshi et al., 2021, p. 6). However, 15 countries only count on one survey on completion rates, which may compromise the accuracy of the ABC estimates. The impact of C-19 might also reveal changes in completion rates when more data becomes available, but not immediately. In addition, in the absence of information on completion at whatever age it occurs, a handful of cases finishing education above 5 years existing expected age for completion (as it is modelled in ABC), might remain unknown to decision-makers. Therefore, in this regard, a critical challenge relates to the generation of data that considers C-19 effects on completion rates and that accounts for late completion of primary and secondary at an adult age.

The second element in this target and indicators has to do with the proportion of children and young people who achieve a minimum proficiency level in reading and mathematics at the end of primary and lower secondary education. Currently, these data sources include the following:

- International assessments (e.g., PISA, PIRLS, TIMSS)
- Regional assessments (e.g., ERCE, SACMEQ, PASEC)
- National assessment data collected through the Catalogue of Learning
- Assessments (CLA) and/or available in national reports
- Population-based assessments:
    - Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA)
    - UNICEF Multiple Indicator Cluster Surveys (MICS)
    - People's Action for Learning (PAL) NETWORK (e.g., Annual Status of Education Report (ASER), UWEZO, etc.) (Adapted from UNESCO Institute for Statistics, 2021, p. 2)

Some challenges have been identified previously, including a need to define minimum proficiency levels for Reading and Mathematics, harmonise different assessments and data sources on educational achievement, and integrate data from early school leavers regarding their proficiency levels (UNESCO Institute for Statistics, 2021). Concerning the first challenge, definitions have been produced and can be seen in Table 2.

*Table 2: Minimum proficiency levels defined by each learning assessment*



| Educational Level | Descriptor |
|---|---|
| **Reading** | |
| **Grade 2** | They read and comprehend most of written words, particularly familiar ones, and extract explicit information from sentences. |
| **Grade 3** | Students read aloud written words accurately and fluently. They understand the overall meaning of sentences and short texts. Students identify the texts' topic |
| **Grades 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. |
| **Grades 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. |
| **Mathematics** | |
| **Grades 2-3** | Students demonstrate skills in number sense and computation, shape recognition and spatial orientation. |
| **Grades 4-6** | Students demonstrate skills in number sense and computation, basic measurement, reading, interpreting, and constructing graphs, spatial orientation, and number patterns. |
| **Grades 8 & 9** | 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.*

## Some of the challenges in measuring SDG4 indicators related to Learning Outcomes and Skills

Despite these definitions being helpful for cross-country comparisons, operationalising the descriptors in an internationally comparable fashion still presents an important set of challenges. Scrutinising these caveats can help researchers and decision-makers identify what measures and monitoring procedures are appropriate for assessing learning outcomes and skills, especially if comparisons are deemed essential. We explore Reading first.

For example, Grade 2 - Reading might require the assistance of an enumerator who verifies a child's capacity to read and comprehend what they read. Although this is not explicit, this descriptor might suggest that reading is to happen aloud, potentially misleading the evaluator's judgement when reading occurs in silence. This aspect of the minimum proficiency level has been rarely discussed; however, it seems necessary to clarify whether the expectation is that reading takes place in a particular manner, and why. Moreover, reading and comprehending are different, conceptually, and pedagogically (reference). It is worth noting that the descriptor also alludes to the child's capacity to extract information. This is yet another element which leads to wondering whether an average of all components in the descriptor is to be produced upon demonstration of a child's capacities, or whether a ponderation would be more appropriate. If the latter is decided, which aspect should be given more importance, and why?

Grade 3 – Reading also might need the presence of an enumerator to verify that reading aloud takes place. For large-scale assessments, this could potentially be done by audio-recording student reading and marking assessments centrally; however, a series of technological and logistical implications arise, which might prevent the evaluation of children's reading, particularly from low-income contexts and remote areas. While accuracy might seem straightforward, its evaluation needs to be sensitive to context given the different accents and variants in pronunciation of the same words within a region or country. Fluency has been subjected to debate as well, particularly regarding its relevance in reading comprehension as opposed to the capacity of a child to decode written language at a rhythm considered appropriate for the age and audience (reference). Like Grade 2 – Reading, these educational level assessments need to agree on a suitable way to assign a marking ponderation to the various components within the descriptor.

Grades 4 & 6 – Reading requires interpretation and explanation, which education specialists might find straightforward to corroborate in children's reading. Still, by mentioning different types of texts in the indicator, the actual texts employed during assessments might vary substantially from context to context, and this may be dictated by curriculum, culture, and history, among other reasons. Although UNESCO recognises text types may include narrative, descriptive, expository, procedural, and verbal interaction, in some places, children in grades 4 & 6 might be reading newspapers and local poetry, while others might focus on tales, biographies, etc., and standardised exams might include some of those or different types of texts. Because of this, evaluating reading proficiency within a given context might be feasible; however, meaningful and fair comparisons across countries might have limitations given the nature of the texts that students read in this age bracket. Evaluators might also consider how to proceed in cases where students are able to provide evidence of their connecting primary ideas from texts with either personal experience or broader knowledge, but not both.

Grades 8 & 9 – Reading carries the caveats from earlier grades and adds a component relating to drawing conclusions. In this regard, the selection of texts for students of these grades needs to make it possible for children to generate conclusions, either from the general text or characters, rather than present them altogether. Should this situation go inadvertent, results on the students' capacity to demonstrate this level of proficiency might be jeopardised.

Mathematics proficiency in Grades 2 – 3 and Grades 4 – 6 are concerned with number sense and computation. Critically, the meaning and pedagogical actioning of number sense is still in debate as it may include several areas related to understanding numbers, including,

- number meaning
- number relationships
- number magnitude,

- operations involving numbers referents for numbers and quantities (The National Council of Teachers, 1989 cited in Gersten & Chard, 1999).

For UNESCO, number sense includes skills such as reading, writing, comparing, and ordering numbers (UNESCO Institute for Statistics, n.d.-a). Yet, grasping a child's number sense can be difficult as it is deemed a capacity individuals develop in a multitude of manners:

Number sense is highly personalized and is related to what ideas about numbers have been established and also on how those ideas were established. Students highly skilled at paper/pencil computations (often the gauge by which success in mathematics is measured) may or may not be developing number sense (Mcintosh et al., 2005, p. 211).

Furthermore, Grades 4 – 6 – Mathematics could provide a better explanation of basic measurement, for example, relating this to specific measures of dimensions, physical properties of objects, etc., which might provide a different picture of the child's proficiency in measurement skills, depending on the context, curriculum, etc. It is worth mentioning that Mathematics proficiency in the indicators Grades 2-3, 4-6, and 8 & 9 include – like the Reading indicator counterparts – several components which currently seem to be equally weighted, and yet, it needs to be explained why and whether ponderations might be appropriate, in line with sound pedagogical reasoning. Finally, although SDG4 4.1 and indicators seek to grasp proficiency in Reading and Mathematics in Grades 2, 3, and at the end of primary and secondary, an additional caveat for assessments is the time during the academic year when these skills are being evaluated. This is important because differences in curriculum programmes might impact the extent to which children from different contexts demonstrate their skills in the different realms this target focuses.

Given all the limitations explained above, 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 Rossetta 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. The proposal of this agenda will be informed by interviews with relevant stakeholders and experts in the international educational measurement community.

## 3. Harmonisation of different assessments and data sources on educational achievement

As mentioned earlier, large-scale assessments of students' academic results in Reading and Mathematics have been employed for assessing learning outcomes and skills, and the UNESCO Institute for Statistics (n.d.-a) has generated illustrative tables showing the specific assessments in line with Grade/age and relevant descriptors. As an example, for Grade 2-3 – Reading and Mathematics, PASEC 2014-2019 and

ERCE 2013-2019 are reference assessments to monitor minimum proficiency levels. An immediate concern arising is whether all UN members participate in these two assessments, and how comparable the data might be with those who rely on other types of examinations, including UNICEF's Multiple Indicator Cluster Survey (MICS). This is particularly important because while MICS takes 15 minutes to complete (Cardoso, 2020; Cardoso et al., 2020), TIMSS, PASEC or ERCE take longer and are more comprehensive in exploring proficiency in reading and mathematics.

## 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 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 the International Association for the Evaluation of Educational Achievement (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 so-called **Policy linking** method (UNESCO Institute for Statistics, 2021). The Policy Linking methodology 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.

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 a 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. The GPF helps to detect gaps/misalignment and provides a global reference for revising standards, curricula, materials, teacher training, and assessments.

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[(*)]: 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.

To produce reliable benchmarks for international reporting, the Policy Linking Toolkit (PLT) specifies five criteria. These include ensuring 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 the 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.

### 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.

## 4. Harmonisation of context questionnaires from different assessments

Beyond the 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 low- and middle-income countries. 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.

### Overview of existing context questionnaires in different assessments and potential uses

A brief overview of some existing context questionnaires used in various international large-scale assessments is as follows:

- *PISA (Programme for International Student Assessment):* PISA includes questionnaires for students, parents and school principals. These questionnaires collect information on students' backgrounds, attitudes and school contexts.
- *TIMSS (Trends in International Mathematics and Science Study):* TIMSS collects data on students' learning environments and teaching practices through questionnaires administered to students and teachers. These questionnaires help to contextualise the assessment results.
- *PIRLS (Progress in International Reading Literacy Study):* PIRLS includes questionnaires for students, parents and school leaders. These questionnaires collect information on students' reading habits, home environment and school resources.
- *ICCS (International Civic and Citizenship Education Study):* ICCS assesses students' civic knowledge and attitudes. It also includes questionnaires for students, teachers and school leaders to understand the context of civic education.
- ERCE (*Regional Comparative and Explanatory Study*): Conducted in Latin America and the Caribbean, ERCE assesses reading and mathematics skills. Contextual questionnaires collect information on students, teachers and schools in the region.
- *PASEC (Programme d'Analyse des Systèmes Educatifs de la CONFEMEN):* PASEC focuses on francophone African countries and assesses student performance in various subjects. Contextual questionnaires are administered to students, teachers and school principals.
- *SAQMEc (Southern and Eastern Africa Consortium for Monitoring Educational Quality):* SAQMEc assesses student performance in literacy and numeracy in countries in southern and eastern Africa. Contextual information is collected from students, teachers and principals.

Harmonising context questionnaires from international large-scale assessments would allow for a more comprehensive and meaningful analysis of educational contexts and outcomes. This data-driven approach would support informed decision-making, policy development, and efforts to enhance the quality and equity of education systems. More relevant for this document, this data harmonisation would allow for disaggregating the data to monitor SDG 4 indicators by different sociodemographic characteristics, such as gender, age, socioeconomic status, parental education, urban/rural, etc.

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.

- *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.

## Challenges in harmonizing these questionnaires

Harmonising context questionnaires in international large-scale assessments presents several challenges, some of which are briefly described below:

- *Cultural and Linguistic Differences*: Variations in cultural and linguistic factors can lead to diverse interpretations of existing questions and responses. Ensuring that questions are culturally sensitive and translate accurately is a significant challenge.
- *Contextual Specificity*: Educational contexts vary widely between countries and regions. Adapting existing context questions to be relevant and applicable across diverse settings can be challenging.
- *Differing Educational Systems*: Differences in educational systems, structures, and policies between countries can complicate the harmonisation of existing context questions that were not originally designed to be universally applicable.
- *Data Collection Methods*: Countries may employ different methods, procedures, and instruments for data collection. Harmonising these methods while maintaining data quality is a complex task when working with existing questionnaires.
- *Response Variability*: Individuals' responses to existing questions may vary based on cultural norms and expectations. Ensuring consistent interpretation and response patterns can be challenging during the harmonisation process.
- *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.
- *Quality Control*: Maintaining data quality and consistency across diverse contexts when harmonising existing questionnaires requires rigorous quality control measures, which can be resource-intensive.

- *Changing Contexts*: Educational contexts evolve over time, and existing context questionnaires may become outdated. Regular updates and revisions are necessary to keep the questionnaires relevant during harmonisation.
- *Political and Cultural Sensitivities*: Some questions in existing questionnaires may touch on sensitive political or cultural issues, making it challenging to find common ground and agree on appropriate wording during harmonisation.
- *Data Standardisation*: Harmonising data formats and coding schemes can be a technical challenge when dealing with existing questionnaires, especially when countries use different systems.

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.

## Proposed strategies to achieve harmonization

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.

## The Need for a Blueprint to Evaluate Assessments for Harmonization

While initiatives like the Rosetta Stone, Policy Linking, and AMPLs have worked to harmonise 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 analyse 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.

## References

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2021). Measuring human capital using global learning data. Nature, February 2020. https://doi.org/10.1038/s41586-021-03323-7

Boissiere, M. (2004). Determinants of Primary Education Outcomes in Developing Countries Background Paper for the Evaluation of the World Bank's Support to Primary Education. www.worldbank.org/oed

Cardoso, M. (2020). School-age learning assessment tools: Frequently Asked Questions on the Foundational Learning Skills module. UNICEF-MICS. https://www.oecd.org/pisa/pisa-for-development/PISAD-Workshop-2020-Manuel-Cardoso.pdf

Cardoso, M., Quintana, E., & Mizunoya, S. (2020). School-age learning assessment tools: The Foundational Learning Skills module Children left behind in learning-measuring foundational learning skills in Multiple Indicator Cluster Surveys (MICS). In PowerPoint presentation. http://uis.unesco.org/apps/visualisations/laci/

Dharamshi, A., Barakat, B., Alkema, L., & Antoninis, M. (2021). Adjusted Bayesian Completion Rates (ABC) Estimation.

Gersten, R., & Chard, D. (1999). Number Sense. Http://Dx.Doi.Org/10.1177/002246699903300102, 33(1), 18–28. https://doi.org/10.1177/002246699903300102

Gust, S., Hanushek, E. A., & Woessmann, L. (2022). Global Universal Basic Skills: Current Deficits and Implications for World Development. http://www.nber.org/papers/w30566

Hanushek, E. A., & Rivkin, S. G. (2006). Chapter 18 Teacher Quality. Handbook of the Economics of Education, 2(06), 1051–1078. https://doi.org/10.1016/S1574-0692(06)02018-6

Hanushek, E. A., & Woessmann, L. (2010). The High Cost of Low Educational Performance. OECD publishing.

Hanushek, Ruhose, & Woessmann. (2016). It pays to improve school quality. Education Next, summer, 52–60.

IEA. (2018). Rosetta Stone Measuring global progress towards SDG4 by linking assessments results to TIMSS and PIRLS International Benchmarks of Achievement. https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/08/GAML6-REF-4-Rosetta-Stone-IEA.pdf

Mcintosh, A., Reys, B. J., & Reys, R. E. (2005). A Proposed Framework for Examining Basic Number Sense. 209–221. https://doi.org/10.4324/9780203990247-23

Patel, D., & Sandefur, J. (2019). A Rosetta Stone for Human Capital (CGD Working Paper).

Sandoval Hernandez, A. (2022). Establishing a concordance between regional (ERCE/PASEC) and international (TIMSS/PIRLS) assessments. http://www.uis.unesco.orgRef:UIS/2022/LO/TD/10

Scheerens, J. (1991). Process indicators of school functioning: A selection based on the research literature on school effectiveness. Studies in Educational Evaluation, 17(2–3), 371–403. https://doi.org/10.1016/S0191-491X(05)80091-4

Scheerens, J. (1997). Conceptual Models and Theory-Embedded Principles on Effective Schooling. School Effectiveness and School Improvement, 8(3), 269–310. https://doi.org/10.1080/0924345970080301

Scheerens, J. (2015). Theories on educational effectiveness and ineffectiveness. School Effectiveness and School Improvement, 26(1), 10–31. https://doi.org/10.1080/09243453.2013.858754

Scheerens, J., Glas, C., & Thomas, S. (2003). Educational Evaluation, Assessment, and Monitoring: A Systemic Approach. Swets & Zeitlinger. https://doi.org/10.1017/CBO9781107415324.004

Torres, R. M. (1999). One decade of Education for All: The challenge ahead.

UN. (2015). Transforming our world: the 2030 Agenda for Sustainable Development.

UN. (2023). Sustainable Development Goals. Website. https://unric.org/en/sdg-4/

UNESCO. (2017). SDG 4 Country Dashboard. https://tcgtest.uis.unesco.org/sdg-4-dashboard/sdg-4-country-dashboard/

UNESCO Institute for Statistics. (n.d.-a). Minimum Proficiency Levels used to report for indicator 4.1.1. Retrieved September 25, 2023, from https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Minimum-Proficiency-Levels-MPLs.pdf

UNESCO Institute for Statistics. (n.d.-b). SDG 4 Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all METADATA. Retrieved September 5, 2023, from https://tcgtest.uis.unesco.org/wp-content/uploads/sites/4/2021/02/Metadata-4.1.0.pdf

UNESCO Institute for Statistics. (2021). 4.1.1 Proportion of children and young people (a) in Grade 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. https://tcgtest.uis.unesco.org/wp-content/uploads/sites/4/2020/09/Metadata-4.1.1.pdf

Unterhalter, E. (2014). Measuring Education for the Millennium Development Goals: Reflections on Targets, Indicators, and a Post-2015 Framework. Https://Doi.Org/10.1080/19452829.2014.880673, 15(2–3), 176–187. https://doi.org/10.1080/19452829.2014.880673

Clarke, M., & Luna-Bazaldua, D. (2021). Primer on large-scale assessments of educational achievement. World Bank Publications. https://doi.org/10.1596/978-1-4648-1659-8