United Nations
Educational, Scientific and
Cultural Organization

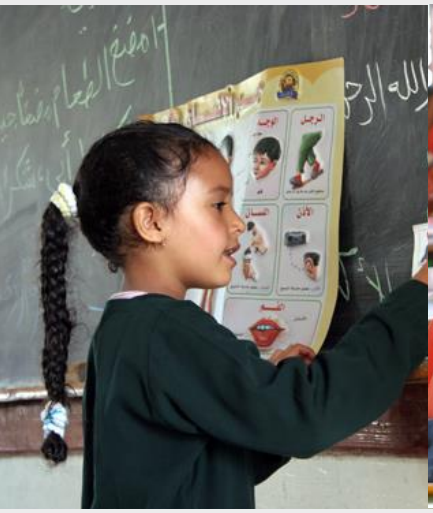
UNESCO INSTITUTE FOR STATISTICS

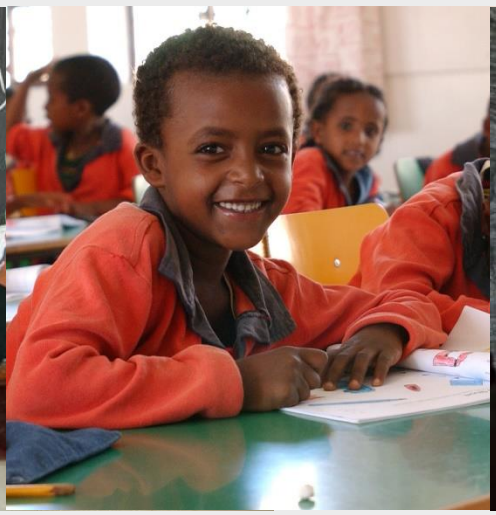
GLOBAL ALLIANCE TO MONITOR LEARNING

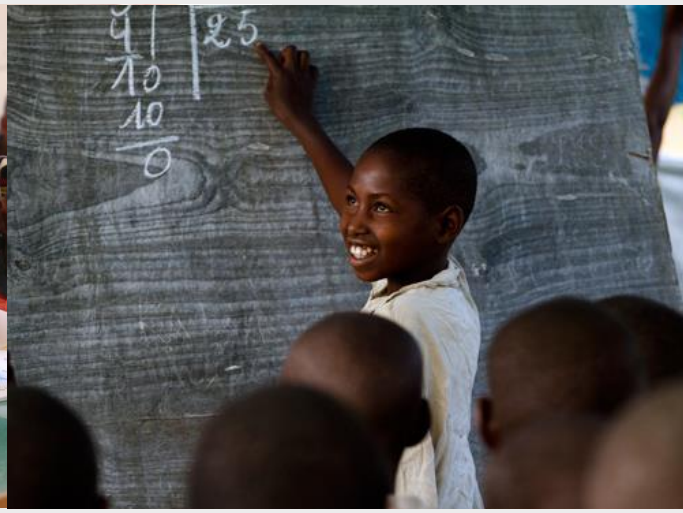
SUSTAINABLE DEVELOPMENT GOALS

Exploring Commonalities and Differences of Regional and International Assessments

GAML Fifth Meeting
17-18 October 2018
Hamburg, Germany







Summary

Background

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

4.1.1 Proportion of children and young people: (a) in 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.

Although indicator 4.1.1 has incommensurable value for our society, it is extremely difficult to measure the progress towards it, for both technical and political reasons. In technical terms, there is an enormous methodological challenge to create globally comparable data to assess the percentage of students achieving a minimal competency level across countries with educational proposals and realities. In political terms, the challenge will come when the leaders need to agree on a "minimum level of competency" in substantive ways.

This paper aims to inform the debate on the possibilities and limitations of developing a global assessment strategy of indicator 4.1.1, considering both the technical and political dimensions of cross-national assessments.

The assessments that were included in this study are: International assessments (e-PIRLS; LANA; PIRLS; PISA 2015; PISA-D; TIMMS); regional assessments (LLECE; PASEC; PILNA; SEA-PLM; SACMEQ), and assessments of foundational skills (ASER; EGMA; EGRA; UWEZO).

Findings of the study

The political challenges of measuring indicator 4.1.1 in a diverse world

- The core of indicator 4.1.1 is measuring the proportion of students at a minimal level of competency. The latent challenge behind this task is to determine what constitutes a minimal level of competency among different education realities worldwide.

- The paper argues that the process of defining the minimal level of competency requires a deep reflection on the practical purpose of SDG4 both for individual countries, especially those that struggle the most, and for the progress in education as a global community.

- The paper also identifies three political challenges involved in defining a minimum level of competency in literacy and numeracy worldwide.

    1. Level of representation of national curriculum in the definition of the minimal level and in the items included in the test.

    2. Expected consequences of the assessment. The results of the assessment of indicator 4.1.1 will be globally disseminated, triggering the political and social consequences of identifying low achieving countries within the international stage. This type of dissemination, although seemingly low-stakes, could create national political pressures for low achieving countries.

3. External or face validity. Countries with high proportions of the population attending school but achieving below the minimal level of competency could react by challenging the validity of the assessment

The technical challenges of defining a minimum international level of competency

It is necessary to establish rigorous, state of the art methodological procedures for international evaluation. The participation of highly reputed institutions is also recommended as it would diminish the threats to external validity of the technical procedure.

- In the first rounds of assessment of indicator 4.1.1, a consortium of specialised international institutions may assess the possibility of establishing some statistical projection or linking of the assessments in order to provide information for the indicator of interest.

- In the mid-term, it would be necessary to define a minimal level of competency in both literacy and numeracy, for each of the three grade ranges. The available information from the different international and regional assessments would be key in order to define the range of competencies to be measured, and especially to consider how to measure skills among low achievers so an assessment provides valuable information for the countries in most need.

Recommendations - Strategies for assessing SDG 4

Strategy 1: use of national assessments to measure SDG4 with adjustments using international assessments. To be implemented in the short-run.

- High levels of external validity for measuring the minimum level of competency established in official curriculum.
- Low levels of international comparability

Strategy 2: equating among international and regional assessments. To be implemented in the medium-run.

- Apparent low cost by using existing assessments.
- Entails performing one equating for each of the grades to be assessed in indicator 4.1.1 and defining new proficiency levels for each scale.
- Technically questionable from a psychometric and substantive point of view.
- Low levels of external validity for representing the national curriculum.

Strategy 3: equating between different international evaluations aiming at similar school grades. To be implemented in the medium or long-run.

- Requires the definition of anchor items that can be shared across the different evaluations and the creation of a consortium of different assessment projects.
- Difficulties of comparison because of the differences in the domains assessed in the different assessments.
- Psychometrically and substantively more robust.
- Low levels of external validity for representing the national curriculum.

Strategy 4: creating a Worldwide Proficiency Assessment on Numeracy and Literacy. To be implemented in the long-run.

- Psychometrically and substantively robust.
- Politically difficult to convince countries to participate in this assessment.

- Requires the participation of technical institutions in the design, implementation, and analysis of test results.
- Low levels of external validity for representing the national curriculum.

1. Introduction[1]

The global leaders of our world set 17 Sustainable Development Goals to improve the quality of people's lives everywhere by 2030. In the context of the knowledge society, achieving the Sustainable Development Goal in education - Goal 4 thereafter SDG4 - seems critical to fulfill the overarching objective of this global agenda. SDG4 aims to promote inclusive and equitable access to quality education as well as to the promotion of development opportunities for all children and youth.

Indicator 4.1.1 operationalizes this goal as the demand to "ensure that all girls and boys complete free, equitable, and quality primary and secondary education leading to relevant and effective learning outcomes" (UNESCO, 2016a, p. 7). The international community agreed upon measuring the progress of this target as the percentage of children and youth achieving a minimal level of competency in literacy and numeracy in three points over time and by sex: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary.

Although indicator 4.1.1 has incommensurable value for our society, it is extremely difficult to measure the progress towards it, for both technical and political reasons. In technical terms, there is an enormous methodological challenge to create globally comparable data to assess the percentage of students achieving a minimal competency level across countries with educational proposals and realities. This entails creating common methodologies to make comparable the data that currently exists as well as promoting the development of new assessments to collect any data that is not yet available. In political terms, the challenge will come when the leaders need to agree on a "minimum level of competency" in substantive ways. Due to the wide variety of educational proposals and realities, curricular logics, and systemic needs of human capital development across countries, the conceptualization of these categories may have significantly different interpretations globally.

Worldwide, there is a wide variety of assessments at regional and international levels that assess education performance for children and youth in literacy and numeracy. Each of these instruments fulfills different purposes and gives relevant evidence for informing decision-making in different educational contexts. In order for global leaders to be able to agree on a common strategy of assessing indicator 4.1.1 and monitoring progress towards it, the information deriving from these assessments needs to be studied to see how comparable it is for defining global indicators. In order to do so, it is necessary to explore the commonalities and differences across regional and international assessments worldwide.

This paper aims to inform the debate on the possibilities and limitations of developing a global assessment strategy of indicator 4.1.1, considering both the technical and political dimensions of cross-national assessments. In doing so, we compare different international, regional assessments, and foundational skills assessments of literacy and numeracy, provide the criteria to make comparison across assessments, address the comparability of all assessments analyzed, and identify the commonalities across assessments to explore the

[1] The paper is commissioned by the UIS and is prepared by Ernesto Treviño and Miguel Órdenes

possibilities of linking assessments to measuring indicator 4.1.1 and recommend strategies about process. Following this analysis, we discuss the implications of agreeing on a common definition of a minimal level of competency at the global scale. Here, we dive into the political intricacies of creating a common definition of the concept of "minimal competency" in reference to a broad variety of education realities between countries. Finally, we suggest four strategies oriented to measure indicator 4.1.1 at a global scale, highlighting the advantages and disadvantages from both a technical and a political perspective.

2. Comparison of assessments

There are numerous regional and international assessments measuring numeracy and literacy in the world. Although many of these assessments aim to measure basic math and language skills, they all show important differences in purpose as well as in concept, methodology, and procedure. Therefore, the study of the differences of cross-national assessments requires careful consideration of multiple criteria of analysis. In this section, we present the assessments under study, the criteria for comparing them, and the results of this comparison.

2.1. Assessments under study

This paper aims to compare a set of international, regional, and foundational skills assessments that measure literacy and numeracy. Particularly, we aim to compare a specific set of assessments in order to inform the decision making process at UNESCO on how to fulfill the measurement requirements of SDG4. In the following table we present the summary of the assessments under analysis:

Table 1. Assessments reviewed

International assessments (non-regional)	Regional assessments	Assessments of foundational skills
• **ePIRLS: Progress in International Reading Literacy Study (online reading)** • **LANA: Literacy and Numeracy Assessment** • **PIRLS: Progress in International Reading Literacy Study** • **PISA 2015: Programme for International Student Assessment**	• **LLECE**: Latin American Laboratory for Assessment of the Quality of Education • **PASEC**: The Programme for the Analysis of Education Systems • **PILNA**: Pacific Islands Literacy and Numeracy Assessment • **SEA-PLM**: The Southeast Asia Primary Learning Metrics	• **ASER**: Annual Status of Education Report • **EGMA**: Early Grade Mathematics Assessment • **EGRA**: Early Grade Reading Assessment • **UWEZO**: Uwezo Annual Learning Assessment

• **PISA D: Programme for International Student Assessment for Development** • **TIMSS: Trends in International Mathematics and Science Study**	• **SACMEQ**: The Southern and Eastern Africa Consortium for Monitoring Educational Quality

At a glance, these assessments have different target populations, contexts, purposes, conceptual focus, methodologies, procedures, and so on. In the next section, we introduce a set of criteria developed to compare these assessments.

2.2. Criteria for comparison: technical dimension

Our starting point is to develop a set of criteria to evaluate the plausibility of comparing the percentage of children and youth achieving a minimal level of competency in literacy and numeracy in three points over time across countries. If the task is to identify a proportion of the population achieving a certain level of competency, then we need to pay attention on how these concepts were measured in different assessments. In order to do so, we need to examine the *design* of the assessments, the *standard setting* procedures to set achievement levels and scores cuts, and the *statistical* procedures used to estimate the distribution of achievement in the population. Differences in these three dimensions may affect significantly the estimations of the proportion of students who achieve the minimum learning proficiency over time.

2.2.1. *Design*: the design of an assessment embodies a set of technical decisions that determine the overarching goal of the assessment, inner rationale of the instrument, and the conceptual framework to be assessed. The design of an assessment defines its purpose, as well as *what* to measure and *how* to measure it. The decisions made in this phase condition the possibilities of what can be done with the data collected. We identified: purpose, population targeted, test construction, domains, potential inferences, sample procedures, and mode of assessing as relevant criteria for comparing assessments' designs. In the following table we define each of these criteria:

Table 2. Design criteria

Criteria	Definition
Purpose	The reason for which the assessment was developed. Two categories fall within this criterion: multipurpose and system-monitoring. Multipurpose assessments can be used for evaluating programmes, defining base lines, diagnose a student population, and so on. System-

	monitoring assessments are oriented to monitor an education system over time.
Target population	The category/categories of students that are to be assessed. This can refer to students at a certain grade level, students of a specific age, or children or youth of a specific or range of ages (not necessarily students).
Test construction	Refers to the nature of what is being assessed: (i) the mastery of a specific content or (ii) the mastery of specific skills. The former will be defined as curriculum-based, the latter as competency-based.
Domains	Describes the specific knowledge and skills evaluated in an assessment.
Inferences	Refers to the validity of the interpretations made from the assessments.
Sample	The subgroup within the targeted population that is included in an assessment evaluation. The rationale to select this subgroup dictates the potential scope of inferences over the targeted population.
Mode of assessment	Refers to how the data was collected. This involves the type of instrument (written or computer-based), the site (school or household), and the administration of the instrument (individual or group).

2.2.2. ***Standard setting*****:** The main purpose of measuring indicator 4.1.1 is to identify a percentage of students performing at a minimal level of competency or achievement. Each assessment uses different approaches of standard setting to build these levels of performance so the scores can be classified in different categories. Standard setting is the procedure of defining frameworks for different performance levels, identifying cut-scores on the score scale defining the threshold between levels, and developing substantive descriptions of what the students classified into any specific level are able to do (Blömeke & Gustafsson, 2017). Particularly, we pay attention to the procedures for identifying cut-scores and how the assessments under study define achievement levels.

Table 3. Standard setting

Criteria	Definition
Cut-score	The knowledge and skills evaluated in a particular assessment.
Levels	The definition of the different categories of performance developed in each assessment.

2.2.3. *Statistical*: The technical treatment of the data collected has to respond to the design of the assessments and, with the interaction between design and statistical analysis, defines the confines of what can be reported on the assessed populations. The technical treatment of the data influences the estimations of the scales, the specific point estimates for countries and

explicit sampling strata, the estimation errors, and, finally, the definition of the performance levels.

Table 4. Statistical criteria

Criteria	Definition
Scaling technique	The statistical methodology used to create the measure of achievement. In general, scaling techniques may include classic test theory, Rasch model, and models with different number of parameters.
Score estimation	The model specification used to estimate individual achievement results. Score estimation may consider either only one estimate per student or the use of several plausible values for each individual in order to describe a distribution of possible scores.
Equating	The procedures to make assessments comparable, either among different forms or over time.

2.3. Addressing the comparability of assessments

The present section compares all the assessments according to the three dimensions introduced in the previous section. For each dimension –design, standard setting, and statistical- we apply each criterion to identify similarities and differences.

2.3.1.Design

a) Purpose: there are important differences in the purposes of the assessments studied. The majority of the assessments aim to monitor achievement in a school system over time and in comparison with other participants (system-monitoring). A small group of assessments - EGMA, EGRA, ASER and UWEZO- aim to conduct system diagnosis or programme evaluation at the national or subnational level (multipurpose). These assessments can be adapted to local needs (national or subnational) and to different measurement challenges. However, these assessments are not designed to generate comparable data at the cross-national level, which may entail important limitations for indicator 4.1.1. For instance, EGMA and EGRA's developers state that those assessments are not designed to make international comparisons but they are made for national diagnosis.

Table 5. Purpose of assessments

Purpose	Assessments
Multipurpose	EGMA, EGRA, ASER, UWEZO
System-monitoring	PILNA, LANA, PIRLS, ePIRLS, TIMSS, SEA-PLM, PASEC, PILNA, SACMEQ, LLECE, PISA, PISA-D

b) Target population: There are two approaches to assess a population: age-based approach (e.g. PISA or ASER) or grade-based approach (e.g. TIMSS). An age-based approach defines a

target population of a specific age or age range. Age-based approaches may be coupled with specifications regarding the specific school grades of the students in the target population. Depending on the purpose of the assessment, some assessments can be used to make comparisons while others cannot. For instance, PISA is an age-based evaluation of different knowledge areas oriented to generate comparable data at the cross-national level and over time. ASER and UWEZO, on the other hand, are age range based assessments and aim to measure the attainment of foundational skills in reading and mathematics in children and youth aged between 5 to 16 years. ASER and UWEZO are not oriented to compare achievement across different locations or over time.

The advantage of an age-based approach is that the evaluation can assess children and youth of similar maturity. However, it can generate biases in relation to the grade levels in which students are enrolled, if any. A grade-based approach assesses students at one or two grade levels. The coverage of the levels varies across assessments. As shown in table 6, the assessments cover almost all grade levels--or ages that can be approximately linked to school grades--in elementary and secondary school. Theoretically, there are four assessments that measure proficiency in grades 2 and 3 (EGMA, EGRA, LLECE and PASEC). However, the designs of EGMA and EGRA do not support comparison across countries or over time. When considering the end of primary (in relation to the International Standardized Classification of Education – ISCED), there are six international evaluations that may be useful to fulfill indicator 4.1.1 (SEA-PLM, LANA, PASEC, PILNA, SACMEQ and LLECE). Furthermore, TIMSS and PISA are the two studies that assess students nearing the end of lower secondary. The advantage of a grade-based approach is that it can assume a similar academic experience among the participants, although it can be quite diverse among countries. However, the ages of students within the grade levels can vary as well (O'Leary, 2001; Ramalingam, 2017), which may affect the inferences over the population.

Table 6. Assessments and grade of evaluation

Grades/Age	Assessments
1st	EGMA, EGRA
2nd	EGMA, EGRA, PASEC
3th	EGMA, EGRA, LLECE
4th	PILNA, LANA , PIRLS, TIMSS
5th	SEA-PLM
6th	LANA, PASEC, PILNA, SACMEQ, LLECE
8th	TIMSS
15 yo (grade 7 or above)	PISA
14-16	PISA-D
5-16 yo	ASER, UWEZO

c) Test construction: For both external validity and comparative purposes, it is important to consider if the assessments are curriculum content-based or competency-based. We established a distinction between these two logics of test construction to identify potential challenges to the external validity due to lack of alignment between regional or international assessments and national curriculums. The definitions of these concepts are not aimed to create rigorous typologies. In fact, making a clear distinction sometimes becomes blurry since the purposes of some assessments aim to reach both. For instance, TIMSS is oriented to assess competencies coupled with some domains of national curricula. The definitions that we establish will be functional to illustrate potential tensions associated to test construction.

In our conceptualization, a curriculum content-based assessment measures the extent to which students know the contents or standards of a particular subject matter. A curriculum competency-based assessment measures the extent to which children can apply competently the knowledge and skills they have learned in their education trajectories. Following these definitions, we classified the assessments studied. For instance, we identified ASER as a competency-based assessment since the purpose is to focus on foundational skills in literacy and numeracy without considering necessarily grade levels. We identified LLECE as a curriculum content-based assessment because the framework of this instrument was developed after studying all the curricula at specific grade levels of the participant countries (see table 7 for more details). Although, there may be overlaps between content and competency-based assessments, there are two challenges that must be considered in relation to these two types of assessments. First, authorities in a country can question the validity of an assessment when the outcomes are below politicians' expectations. Authorities can claim lack of theoretical alignment between the international assessment and the national curriculum, which presumably may impact internal policy dynamics. In some countries the curriculum is mandatory and tied to different policies, such as teacher training and production of textbooks and pedagogical materials. In this case, the results of an evaluation may become high-stakes for the implications on positioning the country in the international arena. Second, curriculum content-based assessments may face important challenges to reach sufficient degrees of comparability in terms of curriculum coverage in the assessments when trying to build an international evaluation based on national curriculum.

Table 7. Competency-based and curriculum content-based test construction

Test construction	Studies
Competency-based	EGMA, EGRA, ePIRLS, PASEC, PILNA, PIRLS, PISA 2015, PISA-D, ASER, UWEZO
Content-based	LANA, LLECE, SACMEQ, SEAPLM, TIMSS

d) Domains[2]: Identifying the actual knowledge and skills assessed is essential to define the commonality among assessments. When we analyzed each assessment framework, we found that domains and subdomains vary widely both in literacy and numeracy and also across grades or ages assessed. For literacy in grades 2 and 3 (EGRA, LLECE and PASEC) the domains and subdomains are radically different, ranging from phonological awareness and reading fluency to text comprehension at different levels. At the end of primary, there is also a wide variety of domains. In LANA, the domains are vocabulary and reading comprehension. In LLECE, the domains are related to the thematic axes of text comprehension, metalinguistic, and theoretical comprehension. LLECE also includes textual interpretation in terms of literary, inferential, and critical comprehension. In PILNA, the domains are reading comprehension and language features. PIRLS focuses on purposes of reading and processes of comprehension. Finally, in SACMEQ, the domains are narrative prose, expository prose, and documents.

In numeracy there is also a wide range of domains measured in the different evaluations. In the case of grades 2 and 3, EGMA measures number identification, quantity discrimination, number patterns, addition and subtraction, and word problems. LLECE focuses on proficiency in numbers, geometry, measurement, statistics and variation. LLECE also includes the cognitive processes of recognition of objects and elements, and solution of simple and complex problems. At the end of primary, the evaluations on numeracy also include a wide range of domains. LANA, for example, focuses on basic facility with numbers, whole number computation, basic fractions, and reading graphs. LLECE, at the sixth grade, evaluates the same domains as third grade such as proficiency in numbers, geometry, measurement, statistics, and variation. Also, it evaluates the cognitive processes of recognition of objects and elements, and solution of simple and complex problems. PILNA covers numbers, operations, measurement and data, time, and money. SACMEQ aims at numbers, measurement and space-data. Finally, TIMSS focuses on content domains (numbers, algebra, geometry, data display and chance) and cognitive domains (knowing, applying and reasoning).

TIMSS and PISA are the two evaluations that may offer information for assessing proficiency at the end of lower secondary, as they focus on students around grades 8 and 9, respectively. In PISA 2015, the domain of mathematical processes includes: the mathematical formulation of situations; the use of mathematical concepts, facts, procedures, and reasoning; and the interpretation, application, and evaluation of mathematical outcomes. Moreover, PISA includes the domains and subdomains of change and relationships, space and shape, uncertainty and data, and quantity, as well as the personal, occupational, societal, and scientific contexts (this is the same for PISA-D). Finally TIMSS, in the case of the end of primary, includes content domains (numbers, algebra, geometry, data display and chance) and the cognitive domains (knowing, applying, and reasoning).

[2] See appendix 1 and 2 for more details.

e) Inferences: Validity is the core concept regarding inferences in the context of measuring indicator 4.1.1. Validity refers to the degree to which the interpretation of test results is empirically and theoretically supported (Plake & Wise, 2014). The validity of inferences takes form during the process of interpreting assessments' results. The intended uses, design, and methods of analysis establish the scope and limits of the type of valid inferences that can be drawn from the assessments.

As can be inferred from the previous paragraph, validity is a key concept when trying to make interpretations about the minimal level of competencies expected for students in different school grades, as established by indicator 4.1.1.

In general terms, the differences in designs, test-construction procedures, domains assessed, statistical methods, target populations, and other characteristics of the different assessments analyzed represent concrete threats to the validity of the inferences on the percentage of students reaching the minimal level of competencies across the different international and regional assessments analyzed. In this regard, it is important to note that there are degrees of validity of inferences. The degree of the validity of inferences depends on the interaction between the claim made from the assessment results and the way in which the design and methods used in the assessment supports such a claim. This means that it may be possible to work towards making assessments more comparable and guarantee greater levels of validity of the inferences on the measurement of indicator 4.1.1. However, the complexity of the task of either trying a statistical linking or an equating among assessments may involve enormous economic, technical, and transaction costs to obtain limited levels of validity in measuring indicator 4.1.1. This strategy should be contrasted with the option of creating a specific assessment for indicator 4.1.1, which may seem cumbersome in terms of effort at first. However, it deserves careful evaluation as this type of initiative will provide the highest possible levels of validity in the measurement. This, of course, under the assumption that both design and methodological procedures are conducted with the maximum scientific rigor.

f) Sample: Sampling design aims at defining the population for which the assessment can make statistical inferences. International and regional studies, in general, use complex samples in order to perform an efficient process of data collection that allows for making inferences at the national level and, also, at explicit strata of the sample--e.g., public and private schools, among others. Traditionally, studies have focused on designing two-stage samples where the first stage of sampling selection are the schools, and the second stage is the selection of students within each school, which may entail either the census of all the students in a specific grade or a random selection of students. It is important to mention that international and regional studies have allowed countries to increase national sample sizes in order to attain data for specific national purposes, such as measuring regions, types of schools, or schools and students participating in policies that need to be monitored. These types of samples aim at measuring the achievement of students that are actually attending school.

In terms of sampling, there is a challenge in measuring student achievement in small countries in terms of population size. It may be the case that a country with less than 20 schools countrywide would show unstable estimates and large standard errors due to the fact that, even performing a census, they may not have sufficient students as to comply with the minimum requirements of the estimation methods available (e.g. this might be the case for some countries that participate in PILNA).

It is important to note that there is a growing scientific and policy interest in measuring achievement of students that are not necessarily attending school. For such a purpose, there are two different approaches in the assessments analyzed. Assessments such as ASER and UWEZO assess children and youth sampling households, thus ensuring to include children and youth inside and outside of the education system. In the other approach, PISA-D is advancing in measuring achievement of both in-school and out-of-school 14-16 years old population by sampling both schools and households using the same test, but different type of administration (group-based and one-to-one, respectively).

Without neglecting the importance of considering the out-of-school population, it is important to recall that indicator 4.1.1 refers to the minimal level of competencies for students attending specific grades in the population. Therefore, in order to focus on the measurement of indicator 4.1.1, it seems necessary to focus on the measurement of achievement for the in-school population. This strategy will fulfill the requirements of SDG4 and diminish the complexity of the measurement. Finally, the measurement of achievement for students in each country can then be complemented with information on the enrollment and attainment rates. Such an exercise can be done, for example, by weighting the score with the net enrollment rate or by graphically showing the relationship between the percentage of students achieving the minimum level of competency and the proportion of the out-of-school population.

g) Mode of assessments: The mode in which the data is collected, the site in which the assessment is conducted, and the way in which the evaluation is administered also may imply important challenges for comparability. EGMA, EGRA, UWEZO, and ASER are collected orally and are administered in a one-on-one fashion. This is because these assessments do not assume that the population evaluated has minimal skills for reading. All other assessments are administered to a group of students. In the case of UWEZO and ASER, the assessments are conducted in the participants' homes because they are meant to capture the skills of the population that is not necessarily included in the education system. UWEZO and ASER differ from the rest of the assessments in that they are exclusively house-based. Another difference among assessments is that some are computer-based while others are paper-based. Taken together, the differences in the mode of assessments may radically change the standardization of the different evaluations (see table 8 for a full summary).

Table 8. Mode of assessment

Assessment	Data collection	Site	Administration
EGMA	Oral	School-based	One-on-one
EGRA	Oral	School-based	One-on-one
ePIRLS	Computer-based	School-based	Group
LANA	Paper-based	N/A	N/A
LLECE	Paper-based	School-based	Group
PASEC	Paper-based	School-based	Group
PILNA	Paper-based	School-based	Group
PIRLS	Paper-based	School-based	Group
PISA 2015	Computer-based	School-based	Group
PISA-D	Paper-based/Computer-based	School-based/House-based	Group/One-on-one
SACMEQ	Paper-based	School-based	Group
SEAPLM	Paper-based	N/A	N/A
TIMSS	Paper, electronic, or online	School-based	Group
UWEZO	Oral	House-based	One-on-one
ASER	Oral	House-based	One-on-one

2.3.2.Standard Setting

a) Cut-scores: Standard setting is a key issue for measuring the concepts included in indicator 4.1.1. In the most basic form, a standard setting method involves three elements: the framework to be assessed, the definition of thresholds to identify different levels of achievement, and the substantive description of each of the levels. One of the main differences among the standard setting methods is the procedure for establishing the cut-scores that mark the separation between one level of achievement from another. The cut-scores procedures will be critical to estimate the percentage of students in a specific level of achievement.

Among the assessments studied, bookmark and scale anchoring were the two main standard setting methods identified. Although they share some features regarding the use of both statistical and substantive information in the definition of achievement levels, the two methods show important differences in the procedures for establishing the cut-scores.

Setting a cut-score through the bookmark involves the work of a panel of experts who 1) list the available scores across all items assessed ordered by the level difficulty; 2) review the items and select the critical items that can signal the cut-scores between one level of achievement from another; 3) discuss these items in several rounds until the panel arrives to an agreement of the cut-scores. These procedures are used by assessments such as LLECE, PILNA, PASEC, and SACMEQ.

Setting a cut-score through scale anchoring follows the following steps 1) the identification of the items that the population evaluated answered correctly; 2) a panel of experts examines the content of each item to determine the kind of knowledge and skills shown by the students who responded correctly to the item; 3) the panel of experts summarizes the list of item competencies in a description of achievement, providing a content-referenced interpretation of the achievement results. This approach is used by assessments such as PIRLS, PISA and TIMSS.

It is important to mention that the rest of the assessments do not provide information on cut-score definition for two different reasons. First, EGMA, EGRA, UWEZO and ASER do not provide information on the setting the cut-scores. It may be the case that such assessments need to be normalized to each of the populations or countries in which they are applied. This is the case for highly recognized test batteries on development around the world—e.g. Woodcock-Johnson and Woodcock-Muñoz or the Wechsler Intelligence Scale for Children (WISC). In such cases, the cut-scores for defining performance levels are established empirically in each context. Second, ePirls, LANA, and SEAPLM do not have information on cut-scores because these assessments or their corresponding reports have not all been completed and, as was mentioned before, the definition of cut-scores requires the analysis of the difficulties of the items. Therefore, it is not feasible to have the definition of cut-scores without having the results of the assessment and the statistics of item functioning.

b) Performance levels: The different procedures for establishing cut-scores will determine the definition of the levels of performance in a substantive fashion. Once the scores are established, a panel of experts describes the different achievement levels. Each achievement level will be associated with a set of knowledge and skills that students at one particular level have a certain probability of mastering. Across the assessments analyzed, we found that the number of levels of achievement varies across studies, ranging from 4 (PIRLS) to 9 (PILNA). This may pose additional challenges to measure minimum proficiency levels at a global scale. Such information requires further and careful analysis, as it will provide a good understanding of the differences across assessments (see table 9 for a synthesis).

Besides the technical specifications, there is an external validity issue regarding what can be considered a minimum proficiency at the global level that carries with it political implications. Furthermore, having assessments that, in the final scale, can accommodate students from populations of widely differing abilities represents a challenge in terms of both measurement and definition of the proficiency levels.

Table 9. Achievement levels in the assessments analyzed

Assessment	Levels
ePIRLS	4 levels of proficiency level with level 4 as the advance level and level 1 as the basic level.
LANA	N/A
LLECE	5 of proficiency with level 4 as the advance level and below 1 as the basic level.
PASEC	Literacy**:** 5 levels (below 1 to 4) with "sufficient" threshold between 2-3 Numeracy**:** 4 levels (below 1 to level 3) with "sufficient" threshold between 1-2
PILNA	9 levels of competency (0 to 8) where 0 is insufficient for both literacy and numeracy
PIRLS	4 levels of proficiency level with level 4 as the advance level and level 1 as the basic level.
PISA 2015	6 levels of proficiency level with level 6 is the advance level and level 1 (and below) is the basic level.
PISA-D	6 levels of proficiency level with level 6 is the advance level and level 1 (and below) is the basic level.
SACMEQ	8 levels of proficiency level with level 8 as the advance level and level 1 as the basic level.
SEAPLM	N/A
TIMSS	4 levels of proficiency level with level 4 as the advance level and level 1 as the basic level.
UWESO	5 levels, with level 1 as the lowest.
ASER	5 levels, with level 1 as the lowest.

2.3.3.Statistical

The statistical methods used in estimating student achievement may have direct consequences on the reported results, even though international and regional evaluations tend to converge in using similar procedures (Cresswell, Schwantner, and Waters, 2015). Here we analyze the different methods in order to assess the comparability of the studies considered in this document.

a) Scaling technique: Even though there are similarities in the scaling techniques used by different studies, the differences among them limit the possibilities of direct comparison. The assessments under analysis provide concrete examples of the challenges of comparability when using different scaling methods. In the case of PISA 2015, the analytical strategy moved from using a one-parameter model to a partial credit model with two parameters. In practice, this change meant that instead of assigning equal weight to all the items as in the one-parameter model, the new scaling models assigned optimal weights based on the ability of students (OECD, 2016). This change was introduced to the trend analysis of PISA and produced

estimates that changed the ranking of the countries in previous rounds of this study. The LLECE studies offer another example. The scaling methods changed between SERCE and TERCE by introducing, in the latter, the method of Plausible Values, which offered more precise estimates of the population by creating a distribution of individual scores (UNESCO-OREALC, 2016). These changes, along with specific differences decisions in the treatment of items and responses within the scaling in SERCE and TERCE, implied that TERCE had to be scaled with the model of SERCE to establish a comparison.

Table 10. Scaling techniques and plausible values

Assessment	Scaling techniques	Plausible values
ePIRLS	N/A	N/A
LANA	N/A	Yes
LLECE	1PL and Classic Theory of Item	N/A
PASEC	1PL	N/A
PILNA	1PL	Yes
PIRLS	2PL, 3PL, Partial Credit Model	Yes
PISA 2015	2PL, 3PL, Partial Credit Model	Yes
PISA-D	N/A	N/A
SACMEQ	1PL	N/A
SEAPLM	N/A	N/A
TIMSS	2PL, 3PL, Partial Credit Model	Yes
UWESO	N/A	No
ASER	N/A	No
EGMA	N/A	No
EGRA	N/A	No

b) Score estimation: The examples of the studies previously described set the basis for the discussion of the score estimation process. In this regard, it is key to mention that the scaling methods are not singular decisions. Instead, under the umbrella of the scaling technique, there is a set of specific decisions. These decisions are related to the treatment of different test formats, the methods for equating, the treatment of non-responses, decisions on the percentage of correct items needed to classify a student under one level of achievement, the criteria for item analysis, and the use of plausible values in order to better reflect the distribution of scores in the population, among others. This set of decisions differs in each study and may influence the estimation of scores and cut-scores for defining the levels of competency.

It is beyond the scope of this document to make an exhaustive comparison of each one of these decisions made in each of the studies analyzed, but a general scan on this issue shows that there are differences in the way studies deal with several of the decisions mentioned above, a situation that introduces difficulties for comparing levels of achievement across studies.

c) Equating: The inclusion of equating as a criterion of analysis is key for two reasons. First, it informs the option of establishing a common scale and achievement levels for different studies. Second, it aims to evaluate the possibilities of measuring indicator 4.1.1 over time. In general terms, equating refers to the methods of establishing a common scale for different test forms. This is generally used in international assessments which use different ways of block rotation, and it is also used to compare achievement results over time. The LLECE studies, PIRLS, PISA, PILNA, SACMEQ, and TIMSS have used equating for different purposes. It is important to remember that equating is not only a statistical procedure, but it requires commonality also in the domains evaluated by the different instruments.

This analysis of the international and regional assessments shows that, as expected, there are no common items across them thus barring attempts at an estimation procedure to link the different assessments. The possibilities of creating a common scale and common achievement levels through a process of equating in the near future poses enormous challenges due to the differences in the content and skills assessed in each of the studies.

Table 11. Summary of criteria in the different assessments

	Design							Standard setting		Statistical	
Assessment	**Purpose [1]**	**Target Population [2]**		**Test Construction [3]**	**Domain [4]**	**Sample [5]**	**Mode [6]**	**Cut-Score [7]**	**Levels [8]**	**Score Estimation [9]**	**Equating [10]**
EGMA	M	1,2,3	Y		N(1, 5)	N/A	OSN	N/A	N/A	N/A	N
EGRA	M	1,2,3	Y		L(5,14)	N/A	OSN	N/A	N/A	N/A	N
ePIRLS	S	4	Y		L(1,4)	N/A	CSG	N/A	4	N/A	N/A
LANA	S	4,6	C		L(1,2)/N(1,4)	N/A	PXX	N/A	N/A	N/A	N/A
LLECE	S	3,6	C		L(3,8)/N(2,8)	R	PSG	BK	5	1PL/CT	Y
PASEC	S	2,6	Y		L(1,3) (1,2)/N(2,8) (3,13)	R	PSG	BK	5/4	1PL	N/A
PILNA	S	4,6	Y		L(3,3) (3,5)/N(3,5)	R/C	PSG	BK	9	1PL	Y
PIRLS	S	4	Y		L(2,6)	R	PSG	SA	4	2PL, 3PL, PC	Y
PISA 2015	S	15*	Y		L(3,10)/N(3,11)	R	CSG	SA	6	2PL, 3PL, PC	Y
PISA-D	S	14-16*	Y		L(3,13)/N(3,11)	N/A	PSG/CHN	N/A	6	N/A	N/A
SACMEQ	S	6	C		L(1,3)/N(1,3)	R	PSG	BK	8	1PL	Y
SEAPLM	S	5	C		L(2,5)/N(3, 10)	N/A	PXX	N/A	N/A	N/A	N/A
TIMSS	S	4, 8	C		N(2,7)	R	PSG	SA	4	N/A	Y
UWEZO	M	6-16*	Y		L(1,5)/N(1,7)	R	OHN	N/A	N/A	N/A	N
ASER	M	6-16*	Y		L(1,4)/N(1,4)	R	OHN	N/A	N/A	N/A	N

[1] Purpose (M: multi-purpose; S: system-monitoring)

[2] Target Population (*age or age range)

[3] Test Construction (Y= competency; C=content)

[4] Domain (L: Literacy; N Numeracy, one parenthesis represents the first grade evaluated, the second parenthesis represents the second grade evaluated). See Appendix 1 and 2 for more details.

[5] Sample (R: representative, C: Census, N/A: Not Available or Not Applicable)

[6] Mode of Assessment (O: Oral, P: Paper-based, C: Computer-based, S: School-site, H: House-based, N: One-on-One, G: Group, X: N/A)

[7] Cut-Score (B: Bookmark; SA: Scale Anchoring; N/A: Not Available or Not Applicable)

[8] Levels of proficiency (amount of level of performance, N/A: Not Available or Not Applicable)

[9] Score Estimation (1PL: 1 parameter; 2PL: 2 Parameters; 3PL: 3 Parameters; CT, Classic Theory of Item; PC: Partial Credit Model)

[10] Equating (N: No, Y: Yes, N/A: Not Available)

3. Commonality across assessments

The comparison of the technical aspects of the international and regional assessments helps to shed light on critical challenges in measuring indicator 4.1.1. In this section, we summarize the lessons learned from comparing the assessments under study according to their design, standard setting, and statistical dimensions.

In terms of design, we identified the purpose of the assessments, the targeted populations, and the domains as three critical aspects for measuring progress according to the indicator 4.1.1. Regarding purposes, we found that the assessments have been developed to fulfill different needs, either to monitor and compare education systems, diagnose an education system, or to evaluate programmes. The assessments with the purpose of monitoring and comparing at a cross-national level are the most appropriate to be considered as robust instruments for measuring the indicator 4.1.1. These exclude assessments such as EGMA, EGRA, ASER, and UWESO, which do not provide systematic and comparable evidence.

Regarding the target population, in order to measure progress according to indicator 4.1.1 at three measurement points (students in grades 2/3, end of primary, and end of secondary), it is necessary to have data that assesses those populations. For the first population (grades 2/3), the comparable assessments available collect minimal data in grades 2/3. PASEC and LLECE are the only assessments that target 2nd and 3rd graders respectively. Therefore, in order to fulfill the requirements of indicator 4.1.1 with the existing assessments, it would be necessary to have a flexible approach to the grades assessed. One approach would be to combine assessments of 2nd (PASEC), 3rd (LLECE), and 4th grades (ePIRLS, PILNA, LANA, PIRLS and TIMSS). For the second population (the end of primary), there are more assessments targeting this time point. LANA, PASEC, PLINA, SACMEQ and LLECE all collect data among 6th graders. In order to increase the coverage of countries for end of primary, SEA-PLM (5th grade) can also be considered within the pool of assessments for covering this second point in time.

For the final target population of lower secondary, TIMSS, PISA and PISA-D would be plausible alternatives. These two assessments have a wide coverage at the international level. However, these assessments may not offer information on the competency of low-performing populations in less developed countries. The tests of these assessments leave significant portions of the populations in developing countries below level 1, meaning that such segments of the population are not able to correctly answer even the less difficult items on the test. Having assessments that cannot identify the minimum competency of vast segments of the populations in developing countries is an essential shortcoming in terms of informing indicator 4.1.1. In this context, PISA-D shows promising opportunities to assess populations of children and youth with low performance. It is important to note that this exercise only scanned the grades assessed and the countries included, but it does not address the crucial point of defining the minimal level of competency for each of the three targeted populations. Furthermore, this analysis of the possible assessments providing information for indicator

4.1.1 should not be taken as suggestions. It is necessary to consider the technical elements analyzed in all the sections of this document regarding the technical possibilities of both establishing a minimum level of competency and measuring it comparatively.

Regarding the domains, the content and skills assessed in each evaluation vary widely. The assessments of content and skills respond to the different needs of the institutions that developed the instruments. This is especially clear for the regional assessments and for the assessments aimed at monitoring the education systems. This lack of overlap generates an important obstacle to find a common ground among the existing assessments for measuring the indicator 4.1.1.

In terms of the standard setting dimension, we found two severe limitations related with the cut-scores definition and the establishment of achievement levels across assessments. In terms of the cut-scores, there are two different methods used to define thresholds of achievement, based on critically different procedures (i.e. experts' agreement versus statistical analysis of the item difficulty), although in the most sophisticated evaluations these methods are complementary (for example, when using the Bookmark method or another procedure that combines statistical analysis of the items with expert agreement). Also, we found important differences in the definition of the levels of competency across assessments. These differences may create obstacles to observing a minimum competency due to the multiplicity of definitions of levels.

In terms of the statistical dimension, we found several similarities in the procedures for analyzing the data and creating comparable scores. Among the comparable assessments, we also found some similarities in statistical techniques for scaling, estimation of scores, and equating. However, there are numerous subtleties within each of these methodological procedures that, while procedurally similar, are not comparable. Furthermore, most of these comparable methods are mainly related to the creation of the scales, but the more pressing issue in measuring the indicator 4.1.1 is having comparable achievement levels.

3.1. The political challenges of measuring indicator 4.1.1 in a diverse world.

The core of indicator 4.1.1 is measuring the proportion of students at a minimal level of competency. The latent challenge behind this task is to determine what constitutes a minimal level of competency among different education realities worldwide. Every country or region in the world has different needs in terms of education development for their citizens. For instance, while some countries are still fighting for increasing coverage in elementary and secondary schools for both girls and boys (e.g. Niger), other countries are focused on keeping the high standards of quality learning for all their citizens (e.g. Finland). These needs are in direct relationship with the level of development of each nation. In terms of public policy, this means that each country will have different expectations of what is the minimal level of knowledge and skills in literacy and numeracy for their citizens. The assessments under study in this report are an expression of this challenge. For instance, when comparing international assessments with regional assessments for similar grades, we can see significant differences

in the domains assessed, both in literacy and numeracy, as stated in previous sections of this document.

In a very diverse world, the task of agreeing upon a minimal level of competency involves not only technical issues as those discussed previously but also political elements. The willingness to create a global goal for improving education across nations is laudable. But the path for measuring that goal faces the international community with the novel problem of defining a globally common measure for all children and youth. In our opinion, the process of defining the minimal level of competency requires a deep reflection on the practical purpose of SDG4 both for individual countries, especially those that struggle the most, and for the progress in education as a global community.

The technical issues for establishing achievement levels have been already discussed. However, the political challenges for defining a minimal level of competency at the international level represent an increased complexity for measuring indicator 4.1.1. The key political challenge comes from the indicator using the word "minimal", which entails that there is some degree of agreement worldwide on what is an acceptable minimum.

There are at least three political challenges involved in defining a minimum level of competency in literacy and numeracy worldwide. The first challenge relates to the level of representation of national curriculum in the definition of the minimal level and in the items included in the test. The vast majority of countries have defined learning objectives through their official curriculum. Therefore, it is expected that when defining a minimal level of competency, countries may be concerned that their curriculum is not sufficiently represented in the assessment nor the definition of the minimal level. Even in countries where a national curriculum is not compulsory, a set benchmark for minimal competency is still expected to guide the teaching and learning activities in the school system.

The second challenge, which is linked to the first, relates to the expected consequences of the assessment. The results of the assessment of indicator 4.1.1 will be globally disseminated, triggering the political and social consequences of identifying low achieving countries within the international stage. This type of dissemination, although seemingly low-stakes, could create national political pressures for low achieving countries.

In that context of political pressures, the third challenge of external or face validity may arise. Countries with high proportions of the population attending school but achieving below the minimal level of competency could react by challenging the validity of the assessment, potentially arguing that such results are due, at least in part, to the low levels of representation of their national curriculum in the test. Questioning the validity of the test, even if the assessment complies with the highest technical and scientific standards, may create doubts about the results within the international community.

The three challenges outlined in this section for assessing the SDG4 may arise because indicator 4.1.1 is labeling an achievement level as "minimal," establishing a standard without international agreement and only partial representation of the myriad of national curricula. This semantic issue of labeling an achievement level of an international assessment as "minimal" is the core political problem that may even involve issues of sovereignty.

Despite the political difficulties outlined here, there is empirical evidence that 133 countries participate in at least one of the comparable regional and international assessments. These evaluation initiatives bring relevant information to countries for monitoring their education systems and, in some cases, participation in these international initiatives is used to build technical capacities at national level. Therefore, the high interest of countries to participate in international evaluations, along with the mandate of United Nations to assess SDG4, are key counter-arguments related to the political issues in defining a minimal level of competency worldwide.

3.2. How to define a minimum international level of competency

In order to diminish the political challenges in the definition of a minimal international level of competency in literacy and numeracy, it is necessary to establish rigorous, state of the art methodological procedures for international evaluation. Currently, there are several institutions--IEA, OECD, ACER, ETS, UIS among others--that may be able to participate in a consortium focused on defining a minimal level of competency for assessing SDG4. The participation of highly reputed institutions would diminish the threats to external validity of the technical procedure.

The technical challenges of defining a minimal level of competency by combining data from the different studies seem to be enormous and the benefit in technical terms, which is key for external validity, may be low. However, in the first rounds of assessment of indicator 4.1.1, a consortium of specialized international institutions may assess the possibility of establishing some statistical projection or linking of the assessments in order to provide information for the indicator of interest.

In the mid-term, it would be necessary to define a minimal level of competency in both literacy and numeracy, for each of the three grade ranges. The fact that the indicator refers to competencies would be an advantage for measurement, since the notion of competencies reduces the pressure of linking the international evaluation to specific contents in the national curriculum. However, this issue would not disappear. For defining this minimal level, it is necessary to follow the technical procedures of assessment studies by defining an evaluation framework, establishing the topics of literacy and numeracy, designing and piloting tests, performing the evaluation, and setting the standards based on both the statistical and substantive analysis of the results. In doing this work, the available information from the different international and regional assessments would be key in order to define the range of competencies to be measured, and especially to consider how to measure skills among low achievers so an assessment provides valuable information for the countries in most need.

Substantively speaking, the definition of the minimal level of competency would require the identification of the concrete knowledge and skills that a student needs to master in order to prepare for participating competently in society. This participation entails, at least, possessing the competencies to exert citizenship in a broader sense, to have access to the labor market to gain economic means, and to have tools to conduct the personal project of life. It is true that literacy and numeracy are necessary but not wholly sufficient elements to

promote these three forms of participation. However, it is important to define what kind of competencies are necessary to ensure this minimal level of participation in contemporary societies.

4. Strategies for assessing SDG 4

The information presented above shows the technical challenges of comparability among the different international assessments analyzed, as well as the political issues that may arise during the process of defining and assessing a minimal level of competency worldwide. Despite these challenges, there is a need for measuring indicator 4.1.1 at a global scale. It is important to state that having a comparable measure of proficiency would require the commitment of the different evaluation projects around the world in order to find ways to either link tests or to create a specific assessment for the purposes of SDG4. The challenge of defining a minimum level of competency, however, involves political challenges for convincing countries about the validity of an international definition of such a minimum standard. In order to fulfill the purpose of measuring indicator 4.1.1, there are at least four strategies for the short, medium, and long run. These strategies are summarized in table 12.

Table 12. Summary of strategies for measuring SDG4

Strategy	Implications
Strategy 1: use of national assessments to measure SDG4 with adjustments using international assessments. To be implemented in the short run	• High levels of external validity for measuring the minimum level of competency established in official curriculum. • Low levels of international comparability
Strategy 2: equating among international and regional assessments. To be implemented in the medium run	• Apparent low cost by using existing assessments. • Entails performing one equating for each of the grades to be assessed in indicator 4.1.1 and defining new proficiency levels for each scale. • Technically questionable from a psychometric and substantive point of view. • Low levels of external validity for representing the national curriculum.
Strategy 3: equating between different international evaluations aiming at similar school grades. To be	• Requires the definition of anchor items that can be shared across the different evaluations and the creation of a consortium of different assessment projects.

implemented in the medium or long run	• Difficulties of comparison because of the differences in the domains assessed in the different assessments. • Psychometrically and substantively more robust. • Low levels of external validity for representing the national curriculum.
Strategy 4: creating a Worldwide Proficiency Assessment on Numeracy and Literacy. To be implemented in the long run.	• Psychometrically and substantively robust. • Politically difficult to convince countries to participate in this assessment. • Requires the participation of technical institutions in the design, implementation, and analysis of test results. • Low levels of external validity for representing the national curriculum.

Strategy 1: in the short run, it may be feasible to use national assessment results in order to measure indicator 4.1.1, and the results from international evaluations may play the role of complementary evidence estimating the distance between the national results and an international measure of achievement. This can be a statistical exercise in which the disparity is calculated between the national percentage of students achieving minimum proficiency and the percentage of students achieving the lower levels on international evaluations, and this calculation is then used to weight national results on proficiency levels. This strategy requires careful analyses of the domains, sub-domains, and performance levels of both national and international assessments. Such a strategy may serve as a way to start addressing, as soon as possible, the need to measure indicator 4.1.1 and provide evidence for countries to take actions to improve proficiency.

Strategy 2: The second strategy consists of creating an equating among international studies. This may mean analyzing the items of the different studies (in similar grades or ages) together, and creating an international scale and new proficiency levels based on the items of the evaluations available. This strategy may require creating a consortium of different institutions leading international evaluations to collaboratively participate in the statistical analyses of the items, creation of scales, and qualitative analyses of items in order to define proficiency levels.

Strategy 3: The third strategy can be the equating between different international evaluations aiming at similar school grades. This strategy requires the definition of anchor items that can be shared across the different evaluations. Given the wide range of abilities found in studies that reach a large number of countries, it is important that anchor items are located along the whole distribution of achievement worldwide. Through this strategy, it would be possible to have comparable scales across studies, although the challenge of defining achievement levels

(the most important input for indicator 4.1.1) may remain. This is because the evaluated domains may differ from study to study, and they may not necessarily be covered within similar definitions of performance levels across different studies. This strategy may partially solve the issue of providing comparable evidence on proficiency, but it requires the political will of the institutions leading international evaluations to incorporate the equating questions.

Strategy 4: The fourth strategy entails creating a Worldwide Proficiency Assessment on Numeracy and Literacy, which can be applied in 2024 and 2028 (nearing the year 2030 in order to have results of two points in time by the deadline for assessment of the Sustainable Development Goals). This strategy requires the participation and political will of different institutions leading international evaluations to form an international consortium to create this evaluation. Technically speaking, having unified global evaluations on achievement for grades 2/3, end of primary, and end of secondary will be the more appropriate way to ensure comparability in both the scales and, more importantly for indicator 4.1.1, in the levels of proficiency.

The strategies outlined above aim at providing different approaches to the difficult task of measuring indicator 4.1.1. All of the strategies have advantages and shortcomings in relation to technical issues and feasibility. It is important to state that there is no feasible strategy that can completely overcome the external validity challenge of representing the level of achievement expected by the national policy of each country. This is an opportunity cost of any international evaluation.

5. Conclusion

The evaluation of progress towards SDG4, especially indicator 4.1.1, requires an advancement in international educational evaluation. First, it is necessary to create a political agreement on establishing a minimal level of competency in literacy and numeracy. Although there is a political agreement of countries for defining such minimal level of competencies—which has been materialized through the countries' support of the SDGs—it is necessary to advance in the technical sphere to define an evaluation design and the processes required to define the minimum competency levels in practice.

The analyses presented above suggest that the most technically appropriate form of assessing indicator 4.1.1 is to develop in the mid-term a specific instrument with a clear definition of the minimal level of competency. This may ensure high levels of comparability of the results and avoid technical critiques. Also, this can ensure higher degrees of external validity when measuring indicator 4.1.1. This is essential because, in the political arena, external validity may mean political legitimacy among the participants of the assessment. Furthermore, the effort of trying to link the different assessments available would be enormous with only limited benefit, in terms of the quality of the measurement. As a consequence, the possibility of designing and applying a specific assessment should be seriously considered.

Finally, in any of the strategies, it is essential to have the support and collaboration of the institutions specialized in international evaluation, potentially in the form of a consortium. Bringing to the table key technical institutions, as well as regional and international assessment initiatives, may also ensure both the technical quality of the assessment of indicator 4.1.1 and its political legitimacy.

Bibliography

ASER. (2017). *Annual status of education report (Rural)* (provisional report). New Delhi: ASER Centre.

Blömeke, S., & Gustafsson, J. (2017). Introduction. In S. Blömeke & J. Gustafsson (Eds.), *Standard setting in education. The nordic countries in an international perspective*. Cham: Springer International Publishing.

Carr-Hill, R. (2015) PISA for development technical strand c: Incorporating out-of-school 15- year-olds in the assessment, OECD Education Working Papers, No. 120, OECD Publishing, Paris.

Cresswell, J., U. Schwantner and Waters, C. (2015). *A Review of international large-scale assessments in education: Assessing component skills and collecting contextual data*. The World Bank, Washington, D.C./OECD Publishing, Paris.

Gove, A., & Wetterberg, A. (Eds.). (2011). The Early Grade Reading Assessment: Applications and interventions to improve basic literacy. Research Triangle Park, NC: RTI Press.

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Capelle, F., et al. (2010). *SACMEQ III Project results: Pupil achievement levels in reading and mathematics.* Paris: Southern and Eastern Africa Consortium for Monitoring Educational Quality.

Justin, S. (2016). *Internationally comparable mathematics scores for fourteen African countries* (Working Paper No. 444). Center for Global Development.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., & Martin, M.O. (Eds.). (2013). PIRLS 2016 assessment framework, 2nd edition. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Martin, M. (2015). *Introducing IEA's LaNA for developing countries*. Presented at the 56th IEA General Assembly, Mexico City, Mexico.

Murimba, S. (2005). The southern and eastern Africa consortium for monitoring educational quality (SACMEQ): Mission, approach and projects. *Prospects*, *35*(1), 75–89.

OECD. (2016). OECD (2016), PISA 2015 results (volume I): Excellence and equity in Education, PISA, OECD Publishing, Paris.

OECD. (2016). *PISA 2015 Assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/9789264255425-en

OECD. (2017a). 2015 Technical report - PISA. Retrieved October 7, 2017, from http://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2017b). Pisa for development assessment and analytical framework preliminary version (Draft). Secretary-General of the OECD. Retrieved from

http://www.oecd.org/pisa/pisaproducts/PISA-D%20Assessment%20and%20Analytical%20Framework%20Preliminary%20Version.pdf

O'Leary, M. (2001). The effects of age-based and grade-based sampling on the relative standing of countries in international comparative studies of student achievement. *British Educational Research Journal*, *27*(2), 187–200.

PASEC. (2015). *Education system performance in francophone Sub-Saharan Africa*. Dakar, Senegal: Conference of Ministers of Education of French-Speaking Countries (CONFEMEN).

PILNA. (2016). *2015 Pacific Islands literacy and numeracy assessment (PILNA). Regional report*. Suva, Fiji Islands: Educational Quality Assessment Program (EQAP).

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for educational and psychological testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12.

Ramalingam, D. (2017). Test design and objectives. In P. Lietz, K. F. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments*. John Wiley & Sons.

Reubens, A. (2009). *Early Grade Mathematics Assessment (EGMA): A conceptual framework based on mathematics skills development in children*. Research Triangle Park, NC: RTI International.

SEA-PLM. (2017). *Southeast Asia primary learning metrics (SEA-PLM)*.

UNESCO. (2016a). *Education for people and planet: Creating sustainable futures for all* (Second edition). Paris: UNESCO.

UNESCO. (2016b) *Proposal to develop global learning metrics how to measure SDG 4.1 draft.* UNESCO Institute for Statistics.

UNESCO-OREALC. (2016). *Reporte técnico. Tercer estudio regional comparativo y explicativo, TERCE.* Santiago, Chile: Oficina Regional de Educación para América Latina y el Caribe (OREALC/UNESCO Santiago).

UNESCO. (2017). *More than one-half of children and adolescents are not learning worldwide* (Fact Sheet No. 46). UNESCO Institute for Statistics.

APPENDIX 1: Domains and Subdomains in Numeracy

Assessment	Grade	Domains and subdomains[1]
EGMA	1,2,3	**1. Literacy:** 1.1 Number Identification, 1.2 Quantity Discrimination, 1.3 Missing number (number patterns), 1.4 Addition and Subtraction (both basic facts in Level 1 and double digit in Level 2), 1.5 Word Problems.
LANA	4,6	**1. Numeracy**: 1.1 Basic facility with numbers, 1.2 Whole number computation, 1.3 Basic fractions, 1.4 Reading graphs
LLECE	3,6	**1. Thematic Axe**s: 1.1 Numerical proficiency, 1.2 Geometric Proficiency, 1.3 Proficiency in measurement, 1.4 Statistical proficiency, 1.5 Proficiency in variation **2. Cognitive Processes**: 2.1 Recognition of objects and elements, 2.2 Solution of simple problems, 2.3 Solution of complex problems.
PILNA	4,6	**1. Numeracy**: 1.1 Numbers, 1.2 Operations, 1.3 Measurement and Data, 1.4 Time, 1.5 Money.
PASEC	2	**1. Arithmeti**c: 1.1 Counting, 1.2 Quantifying and handling quantities of objects, 1.3 Performing operations, 1.4 Completing series of numbers, 1.5 Solving problems. **2. Geometry, Space and measurement**: 2.1 Recognizing geometric shapes, 2.2 Determining spatial location, 2.3 Appraising size
	6	**1. Arithmetic recognizing, applying and solving problems using**: 1.1 Operations, 1.2 Whole numbers, 1.3 Decimal numbers, 1.4 Fractions, 1.5 Percentages, 1.6 Series of numbers and data tables **2. Measurement - recognizing, applying, and solving problems involving the concept of size**: 2.1 Length, 2.2 Mass, 2.3 Capacity, 2.4 Surface Area, 2.5 Perimeter **3. Geometry and Space**: 3.1 Recognition of the prospects of two- or three-dimensional geometric shapes, geometric relations and transformations, 3.2 Orientation in and visualization of space
PISA 2015	15*	**1. Mathematical processes**: 1.1 Formulating Situations Mathematically, 1.2 Employing mathematical concepts, facts, procedures and reasoning, 1.3 Interpreting, applying and evaluating mathematical outcomes. **2. Mathematical Content**: 2.1 Change and relationships, 2.2 Space and shape, 2.3 Quantity, 2.4 Uncertainty and data. **3. Contexts**: 3.1 Personal, 3.2 Occupational, 3.3 Societal, 3.4 Scientific.
PISA-D	14-16*	**1. Mathematical processes**: 1.1 Formulating Situations Mathematically, 1.2 Employing mathematical concepts, facts, procedures and reasoning, 1.3 Interpreting, applying and evaluating mathematical outcomes. **2. Mathematical Content**: 2.1 Change and relationships, 2.2 Space and shape, 2.3 Quantity, 2.4 Uncertainty and data. **3. Contexts**: 3.1 Personal, 3.2 Occupational, 3.3 Societal, 3.4 Scientific.
SEAPLM	5	**1. Context**: 1.1 Personal, 1.2 Local, 1.3 Wider world, 1.4 Intra-mathematical. **2. Process**: 2.1 Translate, 2.2 Apply, 2.3 Interpret and Review. **3. Content**: 3.1 Number and Algebra, 3.2 Measurement and Geometry, 3.3 Chance and Data.
TIMSS	4, 8	**1. Content Domains**: 1.1 Number, 1.2 Algebra, 1.3 Geometric Shapes and Measures/Geometry, 1.4 Data Display/Data and Chance **2. Cognitive Domains**: 2.1 Knowing, 2.2 Applying, 2.3 Reasoning.
SACMEQ	6	**Numeracy:** 1.1 Number, 2.1 Measurement, 2.3 Space-data
UWESO	6-16*	**1. Numeracy**: 1.1 Counting Objects (1-9), 1.2 Number Recognition (11-99), 1.3 Place Value (Ones, tens, hundreds), 1.4 Addition (2- and 3-digit numbers without carrying), 1.5 Subtraction (2- and 3-digit numbers without borrowing), 1.6 Multiplication (1-digit facts), 1.7 Division (1- and 2-digit facts).
ASER	6-16*	**1. Numeracy**: 1.1 Number Recognition (1-9), 1.2 Number Recognition (11-99), 1.3 subtraction (2-digit by 2-digit), 1.4 division (3-digit by 1-digit)

* Age range

[1] Domains are in bold and subdomains are listed after the colon.

APPENDIX 2: Domains and Subdomains in Literacy

Assessment	Grade	Domains[1]
EGRA	1,2,3	**1. Phonological awareness:** 1.1 Initial sound identification, 1.2 initial sound discrimination, 1.3 Segmentation (phoneme or syllables), **2. Alphabet knowledge and decoding**: 2.1 Letter name identification, 2.2 Letter sound identification, 2.3 Syllable identification, 2.4 Familiar word reading, 2.5 Non-word reading, 2.6 Dictation. **3. Vocabulary and Oral Language:** 3.1 Listening Comprehension, 3.2 Vocabulary. **4. Fluency:** 4.1 oral reading fluency with comprehension. **5. Comprehension:** 5.1 reading comprehension, 5.2 Cloze-Maze
LANA	4,6	**1. Vocabulary and Reading Comprehension**: 1.1 Reading for Literary Experience (stories) (2 passages), 1.2 Reading to acquire and use information (2 passages).
LLECE	3, 6	**1. Thematic Axes**: 1.1 Text comprehension, 1.2 Metalinguistic and theoretical. **2. Textual Interpretation**: 2.1 Literary comprehension, 2.2 Inferential comprehension, 2.3 Critical comprehension. **3. Writing**: 3.1 Discursive, 3.2 Textual, 3.3 Readability conventions
PILNA	4	**1. Reading Comprehension**: 1.1. Locate directly stated information in a variety of genres. **2. Language Features**: 2.1 recognize the correct grammatical conventions in the use of capitals for proper nouns and in spelling of blends. **3. Writing**: 3.1 write a coherent text that has a few simple ideas by using common story elements, such as a simple title, and has a beginning but the conclusion may be missing or weak.
	6	**1. Reading Comprehension:** 1.1. Read and critically respond to a variety of texts/genres, 1.2 Connect ideas in the titles and in the sequence of events across the texts. **2. Language Features:** 2.1 Identify common grammatical conventions in the use of verb forms and in spelling of some frequently used two-syllable words. **3. Writing:** 3.1 Structure a story that has a beginning a complication and a conclusion, 3.2 Draw additional details beyond the prompts.
PASEC	2	**1. Language**: 1.1 Listening Comprehension, 1.2 Familiarization with written language reading-decoding, 1.3 Reading comprehension
	4	**1. Reading comprehension**: 1.1 Decoding isolated words and sentences, 1.2 Language
ePIRLS	4	**1. Comprehension Processes in the Context of Online Informational Reading**: 1.1 Focus on and retrieve explicitly stated information, 1.2 Make straightforward inferences, 1.3 Interpret and integrate ideas and information 1.4 Evaluate and critique context and textual elements.
PIRLS	4	**1. Purposes for Reading**: 1.1 Literary Experience, 1.2 Acquire and Use Information. **2. Processes of Comprehension**: 2.1 Focus on and retrieve explicitly stated information, 2.2 Make straightforward inferences, 2.3 Interpret and integrate ideas and information, 2.4 Evaluate and critique context and textual elements.
PISA 2015	15*	**1. Situation**: 1.1 Personal, 1.2 Educational, 1.3 Occupational, 1.4 Public. **2. Text**: 2.1 Text format, 2.2 Text Display Space, 2.3 Text type. **3. Aspect**: 3.1 Access and retrieve, 3.2 Integrate and interpret, 3.3 Reflect and evaluate.
PISA-D	14-16*	**1. Processes (Aspects)**: 1.1 Retrieving information, 1.2 Forming a broad understanding, 1.3 Developing an interpretation, 1.4 Reflecting on and evaluating the content of a text, 1.5 Reflecting on and evaluating the form of a text, 1.6 Literal comprehension **2. Situation**: 2.1 Personal, 2.2 Educational, 2.3 Occupational, 2.4 Public. **3. Text**: 3.1 Text format, 3.2 Text Display Space, 3.3 Text type.
SEAPLM	5	**1. Reading Literacy**: 1.1 Text Format, 1.2 Text Type, 1.3 Process **2. Writing Literacy**: 2.1 Text Type, 2.2 Process
SACMEQ	6	**1. Literacy:** 1.1 Narrative Prose, 1.2 Expository Prose, 1.3. Documents.
UWESO	6-16*	**1. Literacy**: 1.1 Letter/syllable recognition, 1.2 Read words, 1.3 Read a paragraph, 1.4 Read a story, 1.5 Comprehend a short story.
ASER	6-16*	**1. Reading**: 1.1 Alphabet, 1.2 Words, 1.3 Paragraph, 1.4 Story.

* Age range

[1] Domains are in bold and subdomains are listed after the colon.