unesco

Institute for Statistics
Technical Cooperation
Group

April 2022

Monitoring of the Sustainable Development Goals Using Large-Scale International Assessments

Monitoring of the Sustainable Development Goals using Large-Scale International Assessments

A strategy for reporting SDG 4 indicators using data from cross-national assessments

unesco
Institute for Statistics

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946. The Organization currently has 195 Member States and 11 Associate Members.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law, and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfil its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical cooperation to Member States for their development policies and projects; and 5) the exchange of specialized information.

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication. The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

Published in 2022 by:
UNESCO Institute for Statistics
P.O. Box 6128, Succursale Centre-Ville
Montreal, Quebec H3C 3J7
Canada
Tel: +1 514-343-6880
Email: uis.publications@unesco.org
http://www.uis.unesco.org

ISBN: 978-92-9189-291-4

Design an layout by: Tiago Vier

Monitoring of the Sustainable Development Goals using Large-Scale International Assessments

A strategy for reporting SDG 4 indicators using data from cross-national assessments[1]

Abstract. International and regional student learning assessments have a vital role to play in monitoring the Sustainable Development Goals (SDGs). First, they provide internationally comparable measures of minimum learning proficiency needed for SDG 4.1. Second, they provide information about students and schools that many countries' administrative data does not include. This paper provides guidance on how to estimate SDG 4 indicators using data obtained from international and regional student assessments tests based on the large scale assessments aimed to measure reading and mathematics, including the reporting of Indicator 4.1.1 and the selection of reporting sources and methodology to estimate indicators within the SDG 4 framework that are not related to cognitive outcomes (including school infrastructure, teachers' professional development, students' exposure to bullying and whether students are learning in their home language) among others.

1 This document has been prepared by Silvia Montoya and Kevin Mac Donald.

Acronyms

AMPL	Assessment of Minimum Proficiency Level
AMPL-b	Assessment of Minimum Proficiency Level for 4.1.1b
CNA	Cross-National Assessments
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EMIS	Education management information system
ERCE	Regional Comparative and Explanatory Study (Estudio Regional Comparativo y Explicativo)
ESCS	Economic, Social and Cultural Status
GPF	Global Proficiency Framework
IEA	International Association for the Evaluation of Educational Achievement
ICAS	International Competitions and Assessments for Schools
ICCS	International Civic and Citizenship Education Study
ICT	Information and Communications Technology
ILSA	International Large-Scale Assessments
ISCED	International Standard Classification of Education
LLECE	Latin American Laboratory for the Assessment of the Quality of Education (Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación)
MICS	Multiple Indicator Cluster Surveys
MILO	Monitoring Impact on Learning Outcomes
MPL	Minimum Proficiency Level
NLA	National Learning Assessment
OECD	Organisation for Economic Co-operation and Development
PASEC	Programme d'Analyse des Systèmes Educatifs de la CONFEMEN
PIAAC	Programme for the International Assessment of Adult Competencies
PILNA	Pacific Islands Literacy and Numeracy Assessment
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PISA-D	PISA for Development
PLD	Performance-Level Descriptors
PV	Plausible Value
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SDG	Sustainable Development Goal
SEA-PLM	Southeast Asia Primary Learning Metrics
SES	Socio-Economic Status
TIMSS	Trends in International Mathematics and Science Study
UIS	UNESCO Institute for Statistics

Table of Contents

1. Introduction 15
1.1. Objectives 15
2. The Sustainable Development Goal (SDG) 4 cognitive indicators 17
3. Large-scale assessments 19
3.1. Assessment framework and instruments 20
3.2. Information collected by learning assessments programs 20
3.3. Intended population 21
3.4. Sampling 21
3.5. Frequency 21
4. The evaluation of student ability in reading and mathematics in basic education 23
4.1. Interpretation 23
4.2. Methodological definitions needed 23
4.2.1. Definition of the minimum proficiency level (MPL) 24
4.2.2. Methods to link assessment to the Global definition of the MPL 25
4.3. Reporting indicator 4.1.1: operational definitions 29
4.3.1. Assessment used for reporting indicator 4.1.1 29
4.3.2. Mapping of grade to measurement point 30
4.3.3. Mapping of domains to reading or mathematics 30
4.3.4. Alignment to the global MPL 31
4.3.5. Selection of reporting source when various sources are available 31
4.4. How can a country generate comparable data for Indicator 4.1.1? 32
5. Contribution of cross-national assessments (CNAs) to measuring equity 35
6. Contribution of cross-national assessments (CNAs) to measuring SDG 4 non-cognitive indicators 39
6.1. Advantages and disadvantages of using CNAs for SDG indicators 41
7. Methodology for producing SDG 4 indicators based on CNAs 43
7.1. Assessment design: Target populations and sampling 43
7.2. Process for estimating an indicator 44
7.3. Further considerations 48
7.4. How the metadata helps 48
8. References 51

List of Tables

Table 1. SDG 4 targets and indicators related to learning outcomes 17
Table 2. Learning assessment surveys 20
Table 3. Minimum proficiency levels for reading 24
Table 4. Linking strategies by data collection status 29
Table 5. Assessments currently used for reporting 30
Table 6. Reporting source prioritization for Indicator 4.1.1 32
Table 7. Reporting options for countries 33
Table 8. Questions and response mapping for defining urban and rural schools 35
Table 9. Variables used to define high and low socio-economic status of students 36
Table 10. Number of countries with available data by SDG indicator and disaggregation. 2010-2020 38
Table 11. Typical questionnaires and data collected in CNAs 39
Table 12. Potential SDG indicators that could be estimates using CNA data 41
Table 13. CNAs, target populations and methods for computing standard errors 43
Table 14. Questionnaires and mapping of responses for SDG 4.5.2 (percent of students learning in their home language) 53

List of Figures

Figure 1. Summary of linking strategies 29
Figure 2. Availability of disaggregated data in cross-national learning assessments 37
Figure 3. Mapping of potential indicators to be reported 40
Figure 4. PISA 2018 question used for bullying 44
Figure 5. Urban/rural question from TIMSS 2015 45

List of Boxes

Box 1. Assessments for Minimum Proficiency Levels 27
Box 2. Policy Linking 28

1. Introduction

The Education 2030 Framework for Action sets out guidelines for taking action in priority areas of education. This target aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". Education is also related to other targets. For instance, education can be linked to public financing of basic services and policy/legal frameworks that permit educational opportunities and integration of different objectives into national education policies and curricula (UIS, 2016).

The United Nations Secretary General's Synthesis Report (UNSG, 2014) recommended that four levels of monitoring be considered: global, regional, thematic and national. This paper focuses on global-level monitoring, which relies on a more limited and carefully chosen set of indicators to provide an overall view of progress towards the targets of the Sustainable Development Goals (SDGs). Among all indicators provided, the international community must address critical measurement challenges within two main groups of indicators: learning outcomes and education equality. Perhaps the most important challenge is to develop statistical standards to provide high-quality data over time and across countries. The ability to analyse and compare national data across countries and years provides insights for measuring performance, driving policy reform and allocating resources equitably to improve learning among all population groups.

Thematic and sectoral monitoring serve as a framework to track progress on a cross-nationally comparable basis with a wider view of sectoral priorities than the global framework, which captures a more limited perspective through a small set of leading indicators.

The SDG 4–Education 2030 Agenda addresses both learning outcomes and education equality, broadly conceptualized, through a universal agenda with indicators that are relevant for all countries. The targets look at learning throughout the life cycle, from early childhood to adulthood. They also go beyond traditional areas of measurement, such as reading and mathematics, to reflect a comprehensive and integrated view of the skills needed in relation to society and the environment.

Equity is emphasized as there is a risk of focusing on quality without addressing the many aspects related to those on the margins and those who have been left behind. By transforming the way learning is understood in different contexts, we can begin to understand how to better promote policies that will enhance education quality and improve learning consequences among those hardest to reach. These issues provide the lens through which countries will look to assess global progress towards achieving their objectives.

1.1. Objectives

Cross National Assessments (CNA) programmes play a vital role in monitoring SDG 4 goals including indicators related to learning outcomes, indicators related to student experiences, teacher and school characteristics, and equity in these indicators between sub-groups. CNAs offer internationally comparable measures of minimum learning proficiency which is a critical indicator of children's access to learning, but they also offer data on SDG indicators related to students, teacher, and schools as well as equity that, halfway through the 2015-2030 SDG period, many countries lack the capacity to measure independently.

The paper describes the strategy being undertaken by the UNESCO Institute of Statistics (UIS) to use CNAs for monitoring SDG 4. It is structured as follows. Chapter II presents the contribution CNAs are making to monitoring SDG 4 indicators related to learning as well the importance of equity-specific goals. Chapter III describes the major global and regional learning assessments and examines the nature of large-scale assessments, including framework, information sought and methodological aspects.

Evaluation of student ability and skills is the subject of Chapter IV, which introduces the methods for

measuring proficiency, the statistical and non-statistical approaches for harmonizing results to the global standards, and guidelines for countries to determine how to report for Indicator 4.1.1.

Chapter V discusses how indicators can be disaggregated by sub-population for estimating inequities in learning, student experience, and measures of education quality and the use of subpopulation data to inform equity dimensions; and the advantages and disadvantages of CNAs as a reporting tool, in terms of both data collection and methodology. The final chapter presents a step-by-step guide on the methodology for producing SDG 4 indicators based on cross-national education assessments.

2. The Sustainable Development Goal (SDG) 4 cognitive indicators

The SDGs and the Education 2030 Agenda ushered in a new era of ambitions for education. Learning outcomes feature prominently in SDG 4 on education, with five targets and six indicators still requiring data on learning outcomes and skills. This is a shift from previous global education targets, such as those in the Millennium Development Goals, which focused solely on ensuring access, participation and completion in formal primary education and on gender equality in primary, secondary and tertiary education.

The Education 2030 targets underscore the extent to which enrolment and participation are the best means to attain good results and learning outcomes at every age and stage, such as school readiness for young children; academic competencies for children in primary and secondary education; functional literacy and numeracy skills; and skills for work, global citizenship and sustainable development for youth and adults. The framework proposes indicators that enable the measurement and comparison of learning outcomes at all levels of education.

The SDG agenda, beyond Goal 4, calls for an explicit focus on equity, including equity-specific goals (Goal 5 on gender equity and Goal 10 on reducing inequalities). In response, education indicators should aim to capture not just national averages but also the variation across different sections of the population defined by group and individual characteristics, such as sex, wealth, location, ethnicity, language and disability (and combinations of these characteristics). **Table 1** summarizes the learning-related indicators.

Table 1. SDG 4 targets and indicators related to learning outcomes

Target		Indicator
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
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	4.2.1	Proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex
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	4.4.2	Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills
4.6 By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy	4.6.1	Percentage of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex
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	4.7.4	Percentage of students by age group (or education level) showing adequate understanding of issues relating to global citizenship and sustainability
	4.7.5	Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience

Source: UIS

3. Large-scale assessments

Large-scale assessments are designed to describe the achievement of students in a curriculum area in an aggregated form to provide an estimate of the achievement level in the education system as a whole at a particular age or grade level. International large-scale assessments (ILSAs) share the same objective, but their assessment is standardized to be conducted in more than one country, such that that their results can be validly compared. Their design is organized based on a curriculum area, although in some cases (such as the Programme for International Student Assessment [PISA]) they are designed based on a set of cognitive skills (math or reading) that a person should have at a specific age.

Normally, these assessments involve the administration of achievement tests to a sample of students, usually focusing on a particular sector in the system (e.g. Grade 8 in the Trends in International Mathematics and Science Study [TIMSS] and International Competitions and Assessments for Schools [ICAS] or 15-year-old students in PISA). One of the less obvious approaches that learning assessments utilize is to include a set of background/contextual questionnaires that permit a better understanding of the drivers of learning. For instance, school-based assessments generally administer questionnaires inquiring about principal (school head, teacher head), school, teacher, Information and Communications Technology (ICT) coordinator, student, home, curriculum and national context. Such questionnaires collect information regarding the student, family, teacher and school to obtain results disaggregated by sex, age, rural/urban location, socio-economic status (SES), language spoken at home, ethnic group, immigration status, disability, etc.

When related to student achievement, this background information can provide insights into how achievement relates to factors such as family SES, levels of teacher training, teachers' attitudes toward curriculum areas, teacher knowledge, and availability of teaching and learning materials.

To provide statistically valid results in sample-based assessments, a representative sample of schools (usually 150 to 200 schools) is drawn from each country, and a sample of students is randomly drawn from within each of the sampled schools, either by sampling entire classrooms or by sampling students across classrooms. Although the best-known ILSAs feature a number of similarities, there are also some substantial differences that need to be considered when comparing the results for different education systems (see Rocher and Hastedt [2020] for a detailed discussion on this point).

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

There are, however, other organizations conducting or supporting regional assessments, such as UNESCO's Regional Comparative and Explanatory Study (ERCE) in Latin America, the Southeast Asian Ministers of Education Organization and UNICEF's Southeast Asia Primary Learning Metrics (SEA-PLM) in South-East Asia, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) in southern and eastern Africa, the Pacific Islands Literacy and Numeracy Assessment (PILNA) by the Educational Quality and Assessment Programme of the Pacific Community, and the *Programme d'Analyse des Systèmes Educatifs de la CONFEMEN* (PASEC) in Francophone countries in West Africa. Further, participation by a school may be voluntary or may be mandated. When voluntary, nonparticipation of some schools will almost invariably lead to biased results and an inaccurate reflection of achievement levels in the education system. Table 2 summarizes the types and characteristics of large-scale learning assessment surveys.

Table 2. Learning assessment surveys

Assessment	Region	Domain, Area	Grade/Age Group	Cycle every .. Years
Global				
Progress in International Reading Literacy Study (PIRLS)	International	Reading	Fourth-grade	4
Trends in International Mathematics and Science Study (TIMSS)	International	Mathematics and Science	Fourth and eighth grades	4
Programme for International Student Assessment (PISA)	International	Reading, Mathematics and Science	15-year-olds	3
PISA for Development (PISA-D)	International			n/a
Regional				
PERCE/SERCE/TERCE/ERCE	Latin American and the Caribbean	Language (reading and writing) and Mathematics	Third and sixth grade	6
Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ)	Africa (Sub-Saharan)	Literacy and Numeracy	Sixth grade	6
Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC)	Africa (Sub-Saharan), Asia (Eastern and South-eastern)	French and Mathematics	Grade 2 and Grade 6	5
The Southeast Asia Primary Learning Metrics (SEA-PLM)	Asia (Eastern and South-eastern)	Reading, Mathematics, Writing, Global Citizenship	Grade 5	4
Pacific Islands Literacy and Numeracy Assessment (PILNA)	Oceania	Literacy and Numeracy	Grade 4 and Grade 6	3
National assessments				
National assessments				n/a

Source: UIS

3.1. Assessment framework and instruments

Most education assessments are directed at measuring a set of cognitive or non-cognitive outcomes that are important for providing information on the performance of the education system on certain indicators. Similar to other national and international assessments, providing an appropriate assessment framework is extremely important. The assessment framework clarifies in detail what is being assessed, why it is being assessed and how it is being assessed. The assessment framework usually includes two main components: the purposes and definition(s) guiding the assessment and an operationalization of the main concepts, which is then used to elaborate a measurement strategy, design or select the appropriate assessment instruments and guide the interpretation of the findings.

3.2. Information collected by learning assessments programs

All education assessments seek answers to one or more of the following questions:

- How well are students learning in the education system (with reference to general expectations, aims of the curriculum, preparation for further learning or preparation for life)?

- Does evidence indicate particular strengths and weaknesses in students' knowledge and skills?
- Do particular subgroups in the population perform poorly? Do disparities exist, for example, between the achievements of (a) boys and girls, (b) students in urban and rural locations, (c) students from different language or ethnic groups or (d) students in different regions of the country?
- What factors are associated with student achievement? To what extent does achievement vary with characteristics of the learning environment (for example, school resources, teacher preparation and competence and type of school) or with students' home and community circumstances?
- Are government standards being met in the provision of resources (for example, textbooks, teacher qualifications and other quality inputs)?
- Do the achievements of students change over time?

3.3. Intended population

In all large-scale assessments, the population to be assessed should be determined by the aims of the assessment and the corresponding information needs. The aim is to collect the data necessary to produce information that will allow each country to measure and monitor.

Once the aims of the assessment, their operationalization and the target population have been determined, it is not necessary to obtain data for each student in the population as in census-based approaches. The inferences of interest can be obtained instead from a suitably designed high-quality sample of students (Rust, 2014). The sample-based approach has a series of advantages. Factors that favour a sample-based approach include substantially reduced costs in test and questionnaire administration, greater accuracy due to the increased possibility to monitor the quality of implementation, and less time for cleaning and managing data as well as for data analysis and reporting. . Nevertheless, while a sample-based approach provides the means to carry out assessments affordably, considerable attention is required in designing and selecting the samples.

3.4. Sampling

The goal of the large-scale assessments that are discussed in this paper is to make inferences about learning outcomes and associated factors for a target population of students. This extends to estimating results for a wide variety of population subgroups and examining the distribution of the variables measured within and across these subgroups. Given these goals, it is not necessary to obtain data for each student in the target populations. Because these assessments are not intended to determine progression of individual students but rather provide inference about outcomes for a population and subgroups of students, they are generally implemented as sample-based surveys. Sampling offers the potential to greatly reduce the cost and burden of an assessment and increase the scope of topics assessed. However, sampling procedures must be correctly designed and analyses of the resulting data must take into account the sampling design (see Chapter VI for more details). Education assessments use different methods and procedures, and a good review of the most common ones can be found in Rust et al. (2017), and Dumais and Gough (2012).

3.5. Frequency

The frequency of international assessments varies from study to study. PISA, for example, is implemented every three years, TIMSS every four years and ICCS in seven-year cycles. The frequency of the assessment should also be determined by its aims. When the purpose of the assessment is to provide information on the performance of the education system on certain indicators, one should take into account that education systems do not change rapidly. Excessively frequent assessments may fail to register any change and create unnecessary cost.

4. The evaluation of student ability in reading and mathematics in basic education

SDG 4 aims to promote inclusive and equitable access to quality education as well as to the promotion of development opportunities for all children and youth. This goal is operationalized 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" and measured using Indicator 4.1.1:

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

4.1. Interpretation

Each measurement point by grade and subject has a verbal definition of what children should be able to achieve as a minimum. For each point of measurement, a threshold divides students above or below a minimum proficiency level. Each of the measurements for Indicator 4.1.1 is thus reported as follows:

- Above the minimum threshold: the proportion or percentage of students who have achieved at least the minimum proficiency level as defined by each assessment;
- Below the minimum threshold: the proportion or percentage of students who do not achieve the minimum proficiency level as defined by each assessment.

4.2. Methodological definitions needed

The indicator needs the following inputs:

- **Domain:** how reading and math are measured. Reading and mathematics are measured at the national level in numerous ways. There is a key distinction between assessments into are informal, formative, short, or designed by teachers, inspectors and district authorities, and formal, typically summative, longer assessments. These distinctions are important for educators because implementing short, formative assessments to monitor progress can lead to the development of more complete summative assessments.
- **Definition of a minimum proficiency level (MPL):** is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) at a given age/grade
- **Linking:** methodologies to harmonize various data sources, including non-official data sources, by a linking procedure to a common definition of the MPL.
- **Proportion:** estimate of the proportion of children above the threshold of the MPL.
- **Operational definitions for reporting:**
    - what assessment will be taken if various assessment results are available for reporting a give country, year and domain
    - what grade/age will be assigned.
- **Sample:** the sample needs to be representation of population
- **Validation:** from government about acceptance of country data.

4.2.1. Definition of the minimum proficiency level (MPL)

The MPL is the benchmark of basic knowledge in a domain (mathematics, reading, etc.) at a given age/grade measured through learning assessments.

To ensure comparability across learning assessments, as a first step it was aligned the verbal definition of MPL for each domain and level in CNAs was established by conducting an analysis of the performance- level descriptors (PLDs) of cross-national, regional and community-led tests in reading and mathematics. The agreement was found in September 2018 as described in Minimum Proficiency Levels (MPLs): Outcomes of the consensus building meeting.

Table 3 presents, in column 3, what level in each assessment aligns with the global MPL definitions for the domain of reading. Thus, for each assessment, the percentage of children/youth above the MPL is the accumulative percentage of children in or above the level in the assessment reflected in column 3. Column 4, on the other hand, shows the level that each assessment utilizes as the reference level.

For instance, for Grades 4 to 6, upper primary, ERCE utilized Level 2 when reporting for countries participating at the regional level, but for global reporting, the UIS utilizes Level 3. The use of Level 3 would mean, under normal assumptions, that the percentage of children above the MPL would be lower for ERCE for each given country if the global level if taken (level 3) than the ones coming from the level of reference taken as a minimum in the region (Level 2). Similar analysis was applied to mathematics and reported in the consensus meeting document.

Table 3. Minimum proficiency levels for reading

Education level	Descriptor	Assessment PLD that aligns with descriptor	Level that the assessment utilizes for reporting
Grade 2	They read and comprehend most of written words, particularly familiar ones, and extract explicit information from sentences.	PASEC (Grade 2) – Level 3	Level 3
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.	Uwezo – Standard 2 (Story with meaning)	Standard 2 (Story with meaning)
		PASEC 2014 (Grade 2) – Level 4	Level 3
		SERCE/TERCE – Level 2	Level 2
		MICS 6 – Foundational Reading Skills	Foundational Reading Skills
		EGRA – Level 9	Not specified
		ASER – Standard 2 (story)	Standard 2 (story)
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.	SACMEQ (Grade 6) – Level 5	Level 3
		PASEC (Grade 6) – Level 4	Level 3
		PIRLS (Grade 4) – Low	Low
		ERCE (Grade 6) – Level 3	Level 2
		PILNA (Grades 4 & 6) – Level 5	Level 4 (Grade 4) and Level 5 (Grade 5)
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.	PISA/PISA4D – Level 2	Level 2

Source: UIS

The agreement on the verbal definition of the MPL and the identification of the proficiency level aligned with that verbal definition was one vital step toward consensus. Still, a second step is needed that is running the psychometric linking exercises in order to check the alignment. Some of the options the UIS has proposed to link assessments to the global proficiency level are described.

4.2.2. Methods to link assessment to the Global definition of the MPL

Researchers have proposed different approaches to relate test scores on one test/form to another test/form. The linking of either a national or a regional assessment to the global proficiency level definition represented by the MPL requires a methodology to identify the same concepts/definition in the national assessment and across assessments.

In general terms, linking attempts to express different tests that were designed for different purposes using a common scale to allow some degree of comparability. In turn, this allows for fair inferences about the subjects (countries) compared.

The process of making different tests comparable, referred to as "moderation", relies on different methodologies and could be classified as either statistical or non-statistical, and has different "usability" depending on whether the data has been collected[2].

Statistical approaches

Statistical approaches used for linking tests are classified as equating, calibration, projection and moderation. The strength of linking depends on assumptions on the degree of similarity between inferences, constructs, populations and measurement conditions. For instance:

- Item calibration: consists of assessment taken by different individuals whose commonality is the sharing of a common set of items that serve to put them onto the same scale and implies:
    - Use of common items in different assessment programmes;
    - Requests being able to put both test in the same scale by joint calibration of test forms;

Has proven to face many difficulties in implementation, from technical to political, as it requires sharing of all the background information.

Currently one possibility is Assessment of Minimum Proficiency Level (AMPL-b) that could be used to align psychometrically a national or regional assessment to 4.1.1b (Box 1).

- Statistical moderation: based on the same students taking two different tests, called the subject-based approach, relies on modelling ability jointly and utilizes the score distribution of two assessments to construct concordance tables mapping the scores on each test into the other. This methodology:
    - Requires two tests;
    - Requires common individuals taking the two tests;
    - Produces a "concordance table" based on psychometric modelling (Rosetta Stone/UIS). The table is not the reporting scale but facilitates by expressing a larger number of countries in the same scale.
- Equating: putting different tests on a common scale, removing unintended differences in test form difficulties and setting up a common scale. Requirements generally include:
    - Tests should measure the same construct (e.g. latent trait, skill, ability);
    - Tests should have the same level of reliability;

[2] *Recalibration of existing data: a validation strategy. Another alternative linking approach is psychometric recalibration, which relies largely on statistical adjustments (Altinok, 2017) , taking advantage of the fact that some countries, referred to as " doubloon countries", participate in more than one cross-national programme. Using several such overlaps has allowed for the identification of roughly comparable proficiency thresholds. It may serve as a validation but it is unlikely to have political buy-in.*

- Equating transformation for mapping the scores of tests should be the inverse function;
- Test results should not depend on the test form an examinee actually takes;
- The equating function used to link the scores of two tests should be the same regardless of the choice of (sub)population from which it is derived.

Non-statistical approaches

Non-statistical methods of linking involve matching up definitions on each test using subjective judgement. This "moderation" or linking is not an application of the principles of statistical inference but a way to specify rules in order to establish agreement for comparing students. Thus, the method calls for direct judgement about the comparability of performance levels between different assessments onto a reference scale.

Among the possible non-statistical approaches is policy linking. Policy linking (Box 2) is a methodology that can be used to link assessments to the MPL descriptor and to set benchmarks (or cut scores) on learning assessments that have already been implemented, allowing those assessments to be aligned across countries and contexts on that common proficiency-level descriptor. The method allows existing assessments to be used to report against SDG 4.1.1.

This methodology can serve to define (and establish) broad standards for the knowledge and skills that students must achieve, with more focus on those definitions than on the specific items or questions used to assess levels. Given that focus, it is also possible to monitor performance and to help understand the meaning of a minimum level of what students are expected to know and be able to do in relation to grade-appropriate contents. This lies at the heart of curricular definitions in any country, and consensual processes and expert inputs are the way forward. In the case of Indicator 4.1.1, the reference definitions come from the Global Proficiency Framework.

Box 1. Assessments for Minimum Proficiency Levels

The COVID-19 pandemic has disrupted education around the world, forcing countries to respond by providing remote learning, adapting curricula and assessments, and developing ways to support the health and well-being of students, teachers and families.

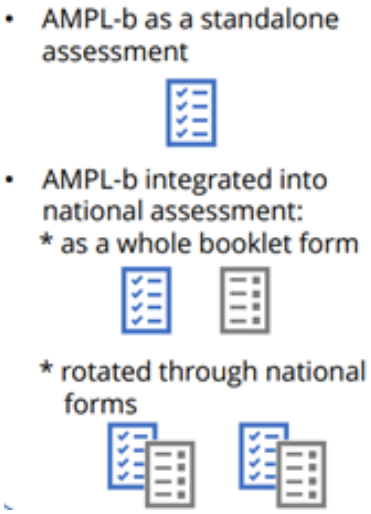
The Monitoring Impact on Learning Outcomes (MILO) project, developed by the UNESCO Institute for Statistics (UIS), is a direct response to that need, providing a way for countries to measure progress against SDG indicator 4.1.1b: the percentage of students at the end of primary school who achieve a minimum proficiency level (MPL) in reading and mathematics. The project aimed to investigate changes in learning outcomes, and the effectiveness of emergency teaching/learning strategies, in six sub-Saharan African countries: Burkina Faso, Burundi, Côte d'Ivoire, Kenya, Senegal and Zambia.

The MILO project developed an Assessments for Minimum Proficiency Levels (AMPLs) toolkit, which was specifically calibrated to the MPL global definition. This toolkit provides an affordable methodological option for countries to measure learning and to develop capacity to regularly generate, analyse and report learning data, as well as monitor progress towards SDG 4. The toolkit has been initially created to monitor indicator 4.1.1b.

The AMPL-b are designed to efficiently measure the proportion of students meeting each MPL. It is important to note that the AMPL-b does not aim to measure the broad range of abilities that children at the end of primary school may exhibit in reading and mathematics. They are targeted at measuring the attainment of a single Minimum Proficiency Level and, on their own, they do not provide further information on where students are at in their learning progression.

AMPL-b could be administered in the future following some of the structures proposed though ideally either of the options that considers implementation along an existing assessment are the preferred choices.

Alternative ways of Using MILO for reporting indicator 4.1.1



Source: UIS

Box 2. Policy Linking

Assessments are not comparable across countries due to different curriculum objectives, coverage of topics, assessment frameworks and items used for national assessments. It also depends on the age at which children enter school and the structure of education, etc. However, the outcomes from assessment that are linked through policy linking might be compared, aggregated, and tracked over time if conditions allows. The policy linking methodology) is method of capacity development and, potentially reporting, that allows countries to map and link existing national reading and/or math assessments to a common scale. This common scale is Global Proficiency Framework (GPF). The methods consists on matching items used in the national assessment to the domains and subdomains as per GPF, and through a benchmarking process might serve to identify the Proficiency Level aligned with the MPL and to estimate, potentially the proportion of students above the MPL as requested by 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).

Policy Linking is implemented through a series of workshops where teachers, subject experts, and assessment and curriculum experts at national and sub-national level based on the experience they gained from the workshops that among others could be listed as:

- Setting benchmarks: The process of setting benchmarks during the workshops includes alignments of items used in national assessment with the Global Proficiency Framework.

- Enhancing capacity: The most important part the workshop is the capacity enhancement of the participating teachers and curriculum and assessment experts by providing opportunities to analyze content and coverage of the items used in the national learning assessment. The process provides international prospective of constructs and sub-constructs through the Global Proficiency Framework, demonstrates the strength and weakness of the used items and helps to develop new relevant quality items for next round of assessments.

- Improving the knowledge of the GPF: A policy linking workshop provides opportunities to assess the closeness of national assessment with the GPF. It encourages the countries to develop items in alignment with the GPF but taking into consideration the national curriculum objectives. It provides guidance to improve theirs learning objectives, curriculum, and assessment framework in future. Many countries have already decided to use the GPF as one of the main reference documents while developing new curriculum and/or new items for learning assessments. The GPF was also helpful to define minimum proficiency levels (MPLs) for next rounds of national assessments.

Countries where Policy Linking had been implemented until 2021

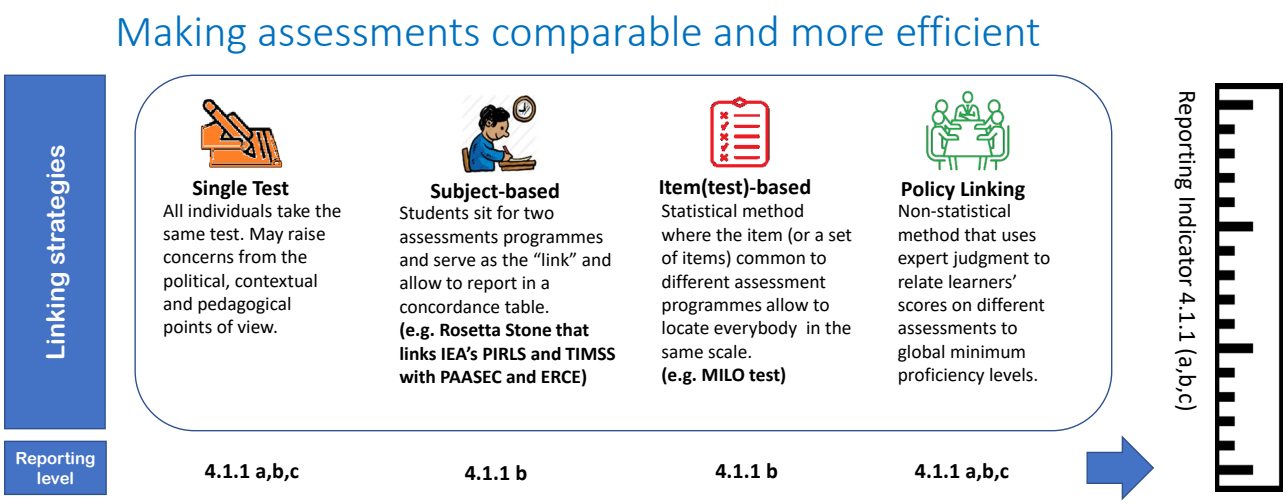
UIS: India, grades 3 and 5 (2019); grade 8 (2021); Bangladesh grades 3 and 5 (2019); Cambodia (grade 6); Lesotho (grade 6) Nepal (grade 5)
USAID: Nigeria (2020), Morocco, Djibouti; Kenya and Senegal;
WBG: Ghana, Gambia
USAID/FCDO/UIS: ICAN/PAL Network

Source: UIS

Summary of options

The choice of linking methodology depends on feasibility, costs and preferences. Figure 1 summarizes the main strategies. The UIS has taken a portfolio approach that includes the two broad sets of possibilities (statistical and non-statistical) defined above. Part of this approach is to assess the degree of the differences that may arise from the different methods.

Figure 1. Summary of linking strategies



Source: UIS

One defining factor in the choice between the different approaches concerns whether the data have already been collected as some of the methods depend on new data collection. Likewise, if data have been collected, some options cannot be implemented as they require a careful design. Table 4 summarizes the options with respect to data collection.

Table 4. Linking strategies by data collection status

	Existing data	New data collection
Subject-based linking (e.g. Rosetta Stone)		Yes
Statistical or item-based linking		Yes
Equating	Yes under certain conditions	
Policy linking	Yes	
Recalibration of existing data	Yes	

Source: UIS

4.3. Reporting indicator 4.1.1: operational definitions

4.3.1. Assessment used for reporting indicator 4.1.1

For global reporting, the UIS currently accepts the assessments listed in Table 5. Both Multiple Indicator Cluster Surveys (MICS) and national assessment are taken under an interim strategy as they need to perform an alignment strategy.

National assessments are used on an increasingly frequent basis but still on an interim basis as policy linking workshops evolve and get implemented.

Table 5. Assessments currently used for reporting

Assessments	Grade 2/3	End of primary	End of lower secondary
ERCE/UNESCO	x	x	
PASEC	x	x	
PILNA	x	x	
PIRLS	x	x	
PISA			x
PISA-D			x
SACMEQ IV		x	
SEA-PLM		x	
TIMSS 4th grade - Math	x	x	
TIMSS 8th grade- Math			x
National assessments to be suitable for reporting need either	Subject to Statistical linking Or sound policy linking On an interim basis if a proficiency level has been identified whose descriptor is aligned with the global MPL		
Modules that measure only one Proficiency Level			
MICS6	x*		
MILO		x	
EGRA.EGMA	Only if the level aligned is administered in the field and subject to representativeness		

Source: UNESCO Institute for Statistics (UIS)

** Interim and subject to alignment with the MPL and assessment of all assessment characteristics*

4.3.2. Mapping of grade to measurement point

1. Grade 2 or 3: Plus one year when primary lasts more than four years according to International Standard Classification of Education (ISCED) levels in the country, except for TIMSS/PIRLS Grade 4, which are mapped to the end of primary education when primary education lasts six years or less.

2. End of primary education: Plus or minus one year from the last year of primary education, except for TIMSS/PIRLS Grade 4, which are mapped to the end of primary education when primary education lasts six years or less, according to the ISCED level mapping in the country.

3. End of lower secondary education: Plus two or minus one year of last grade in lower secondary according to ISCED level mapping in the country.

4.3.3. Mapping of domains to reading or mathematics

1. Reading:

a. If reading is not available for reporting, alternative domains, like language or writing, should be used as long as they cover the domain and subdomains that define the MPL. If there are no alternative domains to align constructs then it is empty;

b. When results are available in different languages, the official or most relevant language in the country is used.

2. Mathematics: alternative domains can also be considered and used for reporting as long as they cover the domain and subdomains that define the MPL.

4.3.4. Alignment to the global MPL

For each assessment, programme alignment should be completed using:

a. CNA: according to agreed alignment from the consensus meeting and its revision

b. National learning assessment (NLA): provided country does not have any CNA. Use statistical linking or policy linking.

4.3.5. Selection of reporting source when various sources are available

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

i. International assessments

ii. Regional assessments

iii. National assessments if they comply with the alignment process.

iv. Population-based assessments if they comply with the alignment process.

Within the specified window of time, the assessment that best maps to the grade is always the preferred choice. Awaiting the linking between assessments, CNA within the window of reporting is the second preferred choice as it allows comparison with other countries. NLA is to be chosen only if no alternative programme for reporting is available and provided alignment had been run through either psychometric or non-psychometric linking.

To illustrate the selection process, we analyse the case of Honduras. Honduras has learning assessment data for Indicator SDG 4.1.1a – Reading in different years, generated from national, regional and international learning assessment programmes:

Year	Source of data
2011	PIRLS 2011 (Grade 4)
2013	ERCE 2013 (Grade 3)
2016	NLA (Grade 3)

Source: UIS

The assessment programme to use for reporting on SDG 4.1.1.a – Reading for Honduras is ERCE 2013. Of the alternatives, PIRLS targets Grade 4, which is one grade above the intended grades of the indicator (Grades 2 and 3). Thus, ERCE and NLA, which both assess Grade 3 students, are better options than PIRLS. However, ERCE prevails over NLA according to the priority of sources of information indicated above as it facilitates comparison between different countries under the same tool (ERCE).

The existence of more than one data point for a given level and domain is another criterion that influences the choice in the interest of tracking progress over time. That restriction is per level and domain; thus, a country could have different reporting sources for a given year for different levels (for instance, ERCE for Grade 2/3 and PISA/TIMSS for end of lower secondary). Accounting for all these caveats, the reporting prioritization is reflected in Table 6.

Table 6. Reporting source prioritization for Indicator 4.1.1

Large Scale Assessment	Sources	Alignment to the global MPL	Priority
International assessments: PISA, PISA4D, TIMSS/PIRLS	Yes	According to consensus meeting	1
Regional assessments: LLECE, PASEC, SACMEQ, PILNA	Yes	According to consensus meeting	2
National learning assessments (NLA)	Yes	If alignment has been done with priority for psychometric alignment	3
Programs aimed to target a proficiency Level			
EGRA/EGMA	Yes	According to consensus meeting if level is pertinent until align-ment is executed, if sample is representative and government agrees.	4*
MICS	Yes	According to consensus meeting on an interim basis until alignment is executed until alignment is executed, if sample is representative.	4*
Citizen Led-Assesments PAL Network (e.g. ASER, UWESO)	Yes	According to consensus meeting on an interim basis until alignment is executed until alignment is executed, if sample is representative and government agrees.	4*

Source: UIS

** Interim reporting until closure of school year 2021*

4.4. How can a country generate comparable data for Indicator 4.1.1?

Currently, a country that wishes to report for Indicator 4.1.1 has the following options:

1) Join a regional or international assessment, if one exists, for the level the country selects to report. Ideally the choice should be consistent with previous data points the country has from previous participation so trends can be estimated.

2) Implement a **national assessment for the first time, ensuring alignment with global reporting.**

a. **Before data collection:** add in the design a booklet that is aligned to global reporting, such as AMPL-b for the end of primary.

b. **After data collection:** run a policy linking workshop for alignment. Pre-data collection activities ideally must include analysis of alignment with the Global Content and Global Proficiency Framework and review the test to ensure sufficient items to identify the cut-off point for the GPF. The rest of the sampling and data procedures should be aligned with international standards.

3) Implement **a national assessment for the second or third time,** ideally in the same grade as the previous time and following the steps below to grant the longitudinal anchoring of the NLA:

a. **Previous rounds:** run a policy linking workshop for alignment to identify alignment of curriculum, test and PLDs as a minimum. Sampling and data procedures need also to

b. **Future data collection:** once the needed adjustment is identified:

i. Add one booklet that is aligned to global reporting, such as AMPL-b for the end of primary, if one exists to allow the linking of the current and previous round;

ii. Run policy linking once data are collected.

Table 7. Reporting options for countries

Level	Early grades	End of primary	End of lower secondary
Global	MICS on an interim basis until alignment is executed	Yes	Yes
Regional	Yes, but only ERCE, PILNA and PASEC for some regions	Yes, but restricted to regional area of the assessment (ERCE/PILNA/ SEA-PLM/PASEC)	
National	Following point 2.b above	In order of priority following point 2.a and then 2.b	Following point 2.b

Source: UIS

5. Contribution of cross-national assessments (CNAs) to measuring equity

The SDG 4–Education 2030 Agenda presents national and international education stakeholders with two important measurement challenges: learning outcomes and education equality. Equity may be emphasized as the need to take into account the many aspects related to those who have been left behind. The SDG agenda includes equity-specific goals (Goal 5 on gender equity and Goal 10 on reducing inequalities).

The most pertinent equity variables are gender, urban or rural location, language spoken at home, immigrant status and SES of pupils. Apart from gender, the remaining equity dimensions do not have a clear analytical definition. Indeed, the distinction between a rural and an urban area is debatable, even within countries. The paragraphs that follow describe the availability of data to define the relevant subpopulations in the major CNAs. This is discussed further in Step 2 below.

Sex of students: Generally, the sex of students is comparable across assessments as being either male or female. The sex of the student is self-reported, although TIMSS also has a sex variable based on administrative records.

Urban and rural location of schools: The location of schools is reported by school directors, and there differences around found in the question asked of the school director. iIn some assessments, the distinction of location of schools is based on the number of people living in the school's area, while in other assessments, the definition is more subjective. For Example, LLECE (Grades 3 and 6) asks: How would you characterize the area where your school is located? and shows the following response options: In an area considered rural (rural), in an area considered urban (urban). PASEC (Grades 2 and 6) propose the following options: Your school is located in… a town (urban), a suburb of a big city (urban), a big village (hundreds of homesteads) (rural), or a small village (dozens of homesteads) (rural). Instead, PISA (15-year-old) has options with more quantitative parameters: Which of the following definitions best describes the community in which your school is located? A village, hamlet or rural area (fewer than 3 000 people) (rural), a small town (3 000 to about 15 000 people) (rural), a town (15 000 to about 100 000 people) (urban), a city (100 000 to about 1 000 000 people) (urban) and a large city (with over 1 000 000 people) (urban). This is discussed in more detail in Step 2 in the methodological section below. Note that assessments collect data on the location type of the school but not on the location type of where each student lives.

Table 8. Questions and response mapping for defining urban and rural schools

Assessment	Population	Question	Responses (mapping)
LLECE 2013	Grades 3 and 6	How would you characterize the area where your school is located?	In an area considered rural (**rural**) In an area considered urban (**urban**)
PASEC 2014/2019	Grades 2 and 6	Your school is located in…	A town (**urban**) A suburb of a big city (**urban**) A big village (hundreds of homesteads) (**rural**) A small village (dozens of homesteads) (**rural**)
PISA 2018	15-year-olds	Which of the following definitions best describes the community in which your school is located?	A village, hamlet or rural area (fewer than 3 000 people) (**rural**) A small town (3 000 to about 15 000 people) (**rural**) A town (15 000 to about 100 000 people) (**urban**) A city (100 000 to about 1 000 000 people) (**urban**) A large city (with over 1 000 000 people) (**urban**)

SEA-PLM 2019	Grade 5	Which of the following characteristics best describes the community in which your school is located?	A village, or rural area (fewer than 3 000 people) (**rural**) A small town (3 000 to about 15 000 people) (**rural**) A town (15 000 to about 100 000 people) (**urban**) A city (100 000 to about 1 000 000 people) (**urban**) A large city (with over 1 000 000 people) (**urban**)
TIMSS 2015	Grades 4 and 8	Which best describes the immediate area in which your school is located?	Urban–Densely populated (**urban**) Suburban–On fringe or outskirts of urban area (**urban**) Medium size city or large town (**urban**) Small town or village (**rural**) Remote rural (**rural**)

Source: UIS

Socio-Economic Status (SES): The term "socio-economic status" generally refers to the relative position of a family or individual on a hierarchical social structure based on their access to, or control over, wealth, prestige and power. In most analyses of the effects of families, schools and other variables on children's academic development, SES is operationally defined with measures describing the occupational prestige, education levels and assets in the children's households. For example, the PISA Economic, Social and Cultural Status (ESCS) index comprises the number of years of education of the parent with the highest education attainment, the occupational status of the parent with the highest occupational status and an index of home possessions that includes measures of wealth, educational possessions and cultural possessions, including the number of books at home. Other assessment programmes offer less data for defining the SES of students. For example, studies conducted by the IEA (i.e. TIMSS and PIRLS) require using a proxy for SES, the "home resources index". Although this index does not capture all dimensions of a basic socio-economic index, it may provide some useful information by distinguishing among students with different socio-economic backgrounds. Another important element is who answers the questions to generate these indexes. For example, in PISA the index is built based on the information from the supplementary questionnaire completed by the students, while in LLECE, the information comes from the supplementary questionnaire completed by the parents of the students.

Table 9. Variables used to define high and low socio-economic status of students

Assessment	Population	Variable	Respondents
LLECE 2013	Grades 3 and 6	Index of the family's socioeconomic status (isecf)	Parents
PASEC 2014	Grade 2	n/a	n/a
PASEC 2014/2019	Grade 6	Socioeconomic index of the student's family (ses)	Students
PISA 2018	15-year-olds	Index of economic, social and cultural status (escs)	Students
SEA-PLM 2019	Grade 5	Socioeconomic status index (ses)	Students and parents
TIMSS 2015	4th grade	Index of home resources for learning (asbghrl)	Students
TIMSS 2015	8th grade	Index of home educational resources (bsbgher)	Students

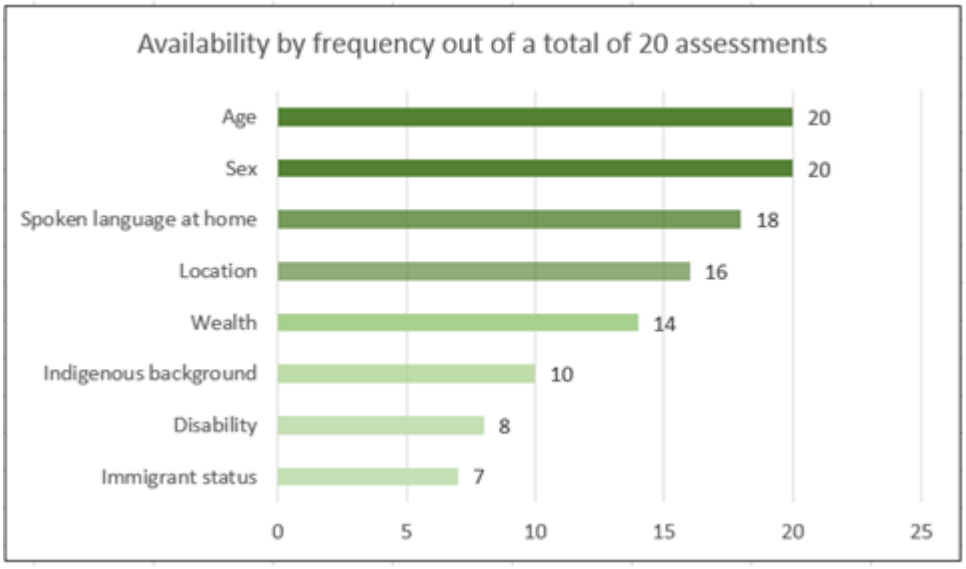
Source: UIS

Language spoken at home, immigrant status and indigenous populations: Another potential equity dimension is the language spoken at home. This distinction is more valid in countries with a large proportion of immigrant populations or when the official language taught in school differs from the language spoken at home. For instance, in some countries, including in European countries, a large proportion of populations with an immigrant background may not speak the official language of instruction. In addition to the language spoken at home, immigrant status may also be used to build a specific equity parity index. Moreover, in many countries, particularly the Americas, differences in educational outcomes between indigenous and other populations are of prime interest for equity. These differences may contribute to specific inequalities in the education system. The assessments also have differences to identify the student's language. For example, PASEC asks directly: What language do you speak at home? PISA asks the question: What language do you speak at home most of the time? (Please select an answer). TIMSS asks: *How often do you speak <test language> at home?.* For more detail see Table 1 in Appendix 1.

Figure 2 and **Table 10** summarize the disaggregation currently available in cross-national learning assessments. Disaggregation by age, sex, home language, location, SES, indigenous background, immigrant status and disability are found in the existing learning assessments, with age and sex included in all the assessments examined. Ethnic background, immigrant status and disability are found to be the least available from current learning assessments. Questions and the ways they are framed put some restrictions on use and comparability, and in some cases harmonization efforts are required for the following reasons:

- Question are collected from multiple instruments in the same assessment (e.g. school head, teacher, student, family)
- Question can vary over time in the same instrument
- Question varies across different assessments
- Question varies between instruments used at country level (e.g. PAL Network ASER and UWEZO ask different questions on school services availability).

Figure 2. Availability of disaggregated data in cross-national learning assessments



Source: UIS

Table 10. Number of countries with available data by SDG indicator and disaggregation. 2010-2020

SDG Indicator	Disaggregation					Total, countries
	Sex	Socioeconomic status (SES) of the students or the students' family	Language of the test at home of students	Immigrant status of students	Location of school	
4.1.0	73					109
4.1.1	97	90	88	90	86	150
4.2.1	76					78
4.2.3	80	70			76	81
4.4.1	97					95
4.4.2	31	31				31
4.5.2	86	86			83	126
4.6.1	41	37		37		50
4.7.4	23	23			21	23
4.7.5	27	29			27	29
4.7.6						3
4.a.2	131	84		28		136
4.c.7	69					87

Source: UIS/UNESCO

6. Contribution of cross-national assessments (CNAs) to measuring SDG 4 non-cognitive indicators

While the primary objective of CNAs is to estimate measures of learning outcomes for a country, they also collect a rich set of background information about teachers, students and schools. From students, CNAs typically collect information about their experience at school, their attitudes towards subjects being taught, and the characteristics of their parents and households in addition to the core demographic information of age and sex (Table 11). From teachers, CNAs collect information about attitudes towards teaching, opinions about teaching resources, education background and ongoing professional development; from schools, CNAs collect information about infrastructure, location and opinions from the school directors about the availability of resources at the school and how they interact with parents. For schools, there is some variation in how data are collected; for example, in TIMSS and PISA, there are few objective questions about infrastructure at the school except for the availability of computers, while in PASEC and assessment by the Latin American Laboratory for the Assessment of the Quality of Education (LLECE), school directors are asked about whether the school has specific types of infrastructure, including toilets and water.

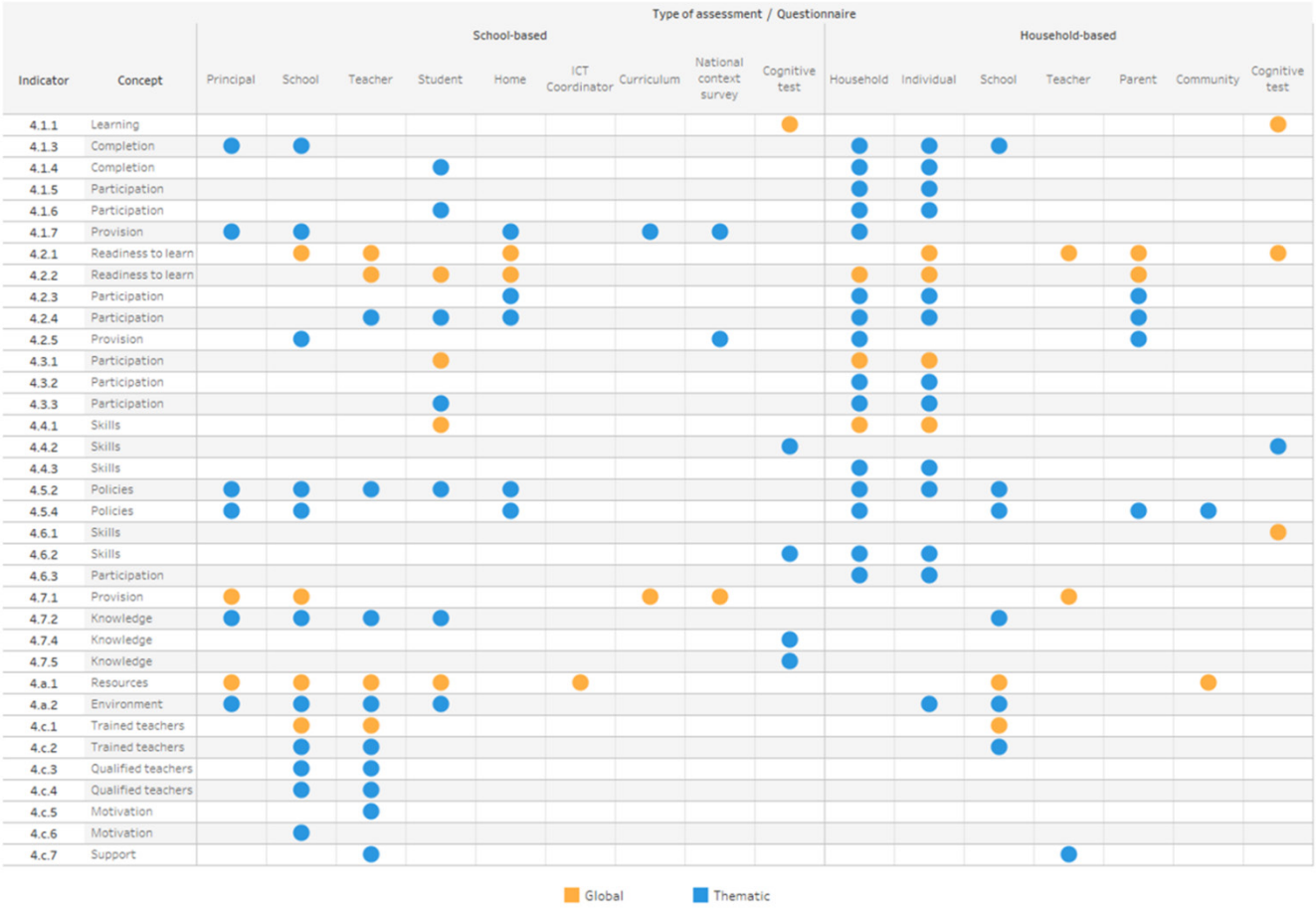
Table 11. Typical questionnaires and data collected in CNAs

Student assessment module	test items (questions) for measuring learning outcomes
Student questionnaire	basic demographic information (sex, age) household and socio-economic background (parents' education, household possessions, language spoken at home) school-related experiences (including exposure to bullying) learning-related experiences (classroom activities) self-perceptions, interests and aspirations related to different subjects (e.g. interest in science, importance of science, whether he or she feels good in a particular subject) use and proficiency of ICT
Teacher questionnaire	demographic and background information (sex, age, years teaching, subjects taught) qualifications and training (education background and qualifications, in-service training and professional development, both frequency and types) types of teaching practices used and challenges faced
School director questionnaire	demographic and background information (sex, age, years of experience) qualifications and education school characteristics (location, availability of resources or infrastructure, public/private) opinions about availability and adequacy of resources management and governance interaction with parents and school communities challenges faced in teaching

Source: UIS

These questionnaires determine whether an SDG indicator can be estimated. They also determine which subpopulations the indicators can be estimated for and how to inform the equity dimension. Generally, these subpopulations include those covering the urban and rural location of the school, the SES of the student (relative to other students, not the population) and sex of student (or teacher). Figure 3 summarizes the indicators that are possible to measure.

Figure 3. Mapping of potential indicators to be reported



Source: UIS

Table 12 presents a list of SDG 4 indicators that are currently being estimated and reported using CNAs and those indicators for which CNAs could be a potential source. In addition to measuring learning achievement (e.g. 4.1.1 and its associated parity indices under 4.5.1), indicators for the proportion of students learning in their home language (4.5.2), the availability of services and infrastructure at schools (4.a.1), the percent of children who have been exposed to bullying (4.a.2) and the percent of teachers who have received professional development in the past 12 months (4.c.7) are currently being reported. In addition to these, there are three broad types of indicators that CNAs may be able to estimate. The first relate to measures of skills that are not the primary objective for measurement for a typical CNA. For example, the percentage of students in lower secondary school showing adequate understanding of issues relating global citizenship and sustainability (4.7.4) could be measured by student questionnaire items about students' understanding of various global issues in PISA. The percentage of students in lower secondary school showing proficiency in knowledge of environmental science and geoscience (4.7.5) could be measured based on scores for the TIMSS science subdomain of Earth Sciences or TIMSS and PISA minimum proficiency in science as a proxy.

A second set of indicators relates to teacher qualifications (4.c). Currently, the percent of teachers with recent professional development (4.c.7) is being estimated using CNAs, but the 4.c indicators around teacher qualifications could also be estimated either using national definitions of "qualified" in conjunction with CNA data or by defining a global definition of "qualified" that could be applied to all countries, for example, those teachers holding a university-level degree. The third set of indicators that CNAs have the potential to estimate includes those related to participation and parity ratios when CNA data is used in conjunction with population data or enrolment rates. For example, the gender parity ratio in enrolment

could be estimated using the ratio of females to males from a CNA divided by the ratio of females to males in the population. Generally, CNAs do not have a comparative advantage for these types of indicators as compared with administrative data on enrolments and household survey data, and these have been excluded from Table 9 but remain a possibility.

Table 12. Potential SDG indicators that could be estimates using CNA data

List of SDG 4 indicators that can be sourced from each assessment:

Type of assessment	Assessment	
School-based	TIMSS	4.1.1 4.2.1 4.2.2 4.a.1 4.c.1 4.1.3 4.1.6 4.1.7 4.2.3 4.2.4 4.5.2 4.a.2 4.c.2 4.c.3 4.c.4 4.c.7
	PASEC	4.1.1 4.2.2 4.a.1 4.c.1 4.1.6 4.1.7 4.2.4 4.5.2 4.5.4 4.a.2 4.c.2 4.c.3 4.c.4 4.c.5 4.c.7
	PIRLS	4.1.1 4.2.1 4.2.2 4.a.1 4.c.1 4.1.3 4.1.7 4.2.3 4.2.4 4.5.2 4.a.2 4.c.2 4.c.3 4.c.4 4.c.7
	SACMEQ	4.1.1 4.2.2 4.a.1 4.c.1 4.1.3 4.1.6 4.2.4 4.5.2 4.7.2 4.a.2 4.c.2 4.c.3 4.c.4 4.c.6 4.c.7
	PISA	4.1.1 4.2.2 4.a.1 4.c.1 4.1.6 4.1.7 4.2.4 4.5.4 4.7.5 4.a.2 4.c.2 4.c.3 4.c.4 4.c.7
	TERCE	4.1.1 4.2.2 4.a.1 4.c.1 4.1.6 4.2.4 4.5.2 4.a.2 4.c.2 4.c.3 4.c.4 4.c.7
	ICCS	4.7.1 4.a.1 4.c.1 4.1.3 4.1.7 4.2.5 4.7.4 4.a.2 4.c.2 4.c.3 4.c.4
	EGMA/EGRA	4.1.1 4.2.2 4.a.1 4.c.1 4.2.4 4.5.2 4.c.2 4.c.3 4.c.4 4.c.7
	ICILS	4.4.1 4.a.1 4.c.1 4.1.3 4.4.2 4.c.2 4.c.3 4.c.4 4.c.7
	EDI	4.2.1 4.2.2 4.2.4
Household-based	Young lives	4.2.1 4.2.2 4.3.1 4.a.1 4.1.3 4.1.4 4.1.5 4.1.6 4.1.7 4.2.3 4.2.4 4.2.5 4.3.2 4.3.3 4.4.3 4.5.2 4.5.4 4.a.2 4.c.7
	MICS	4.2.1 4.2.2 4.3.1 4.4.1 4.1.3 4.1.4 4.1.5 4.1.6 4.1.7 4.2.3 4.2.4 4.2.5 4.3.2 4.3.3 4.4.3 4.5.2 4.6.2
	PAL Network	4.1.1 4.2.2 4.3.1 4.a.1 4.c.1 4.1.3 4.1.4 4.1.5 4.1.6 4.1.7 4.2.4 4.5.2 4.5.4 4.6.2 4.c.2
	STEP	4.2.2 4.3.1 4.4.1 4.6.1 4.1.3 4.1.4 4.1.5 4.1.6 4.2.4 4.3.2 4.3.3 4.4.3 4.5.2 4.6.2 4.6.3
	PIAAC	4.3.1 4.4.1 4.6.1 4.1.4 4.1.5 4.3.2 4.3.3 4.4.3 4.6.2
	EAP ECD Scales	4.2.1 4.2.2 4.2.3 4.2.4 4.2.5 4.5.4
	EHCI	4.2.1 4.2.2 4.2.3 4.2.4
	IDELA	4.2.1 4.2.2 4.2.3 4.2.4
	MELQO	4.2.1 4.2.2 4.2.3 4.2.4
	ITU	4.4.1 4.4.2

Source: UIS

6.1. Advantages and disadvantages of using CNAs for SDG indicators

Using CNAs offers certain advantages and disadvantages compared to using administrative data collected by countries, typically through Education management information system (EMIS) systems.

Firstly, CNAs are generally nationally representative, meaning they include private schools, which are often excluded from countries' EMIS systems.

Secondly, they contain rich data about students' background characteristics, including SES, and therefore permit disaggregation of the estimated SDG indicator, which is typically not possible with administrative data.

Finally, the rich questionnaire data obtained from students, teachers and school directors are often not collected in EMIS systems, including students' exposure to bullying, students' languages spoken at home, teachers' professional development, and schools' characteristics.

The chief limitations of using CNA estimates of SDG indicators compared to indicators derived from administrative data are that (1) they are sample based and subject to sampling error and (2) they are typically implemented for one or two grades (or an age group) rather than all students, which requires a process to map the grade or age group to the education levels described in the definition of the SDG indicator; however, this latter point does not apply to school-level indicators, which rely on data from schools rather than a specific grade. Note that all sample-based sources of indicators (e.g. household survey data or other types of school surveys) would be subject to sampling error, and administrative data may also be subject to error as well. Another disadvantage is that the questionnaire may result in an indicator that does not exactly correspond to the definition used in the SDG; for example, the CNA questionnaire may ask about bullying in the past two years, while the SDG indicator may define bullying in the past year.

7. Methodology for producing SDG 4 indicators based on CNAs

7.1. Assessment design: Target populations and sampling

CNAs are quite similar in their sampling design. Most, including TIMSS, PASEC and ERCE, use students at a specific grade level as the target population, and the sample design is conducted in two stages: first, schools that have students in the target grade are randomly selected, and next, classes are sampled. There is some variation in this second step, particularly when the students at the grade level attend different classes for different subjects. For PISA, the target population is 15-year-olds, so the second-stage sampling is not of classes but rather a random selection of 15-year-olds in the school. For the analyst attempting to estimate an SDG indicator from a CNA, the sample design determines how to estimate an indicator. The two-stage sampling design creates intracluster correlation that must be accounted for in the method for estimating the standard errors and, along with any stratification, unequal selection probabilities of students, which require the use of sample weights.

These sampling issues are well known, and the major CNAs have detailed manuals specifying how to estimate statistics based on their sampling design. They typically rely on a re-sampling method (either jackknife repeated replication or balanced repeated replication) and provide the variables (either the replicate weights or variables used to generate these weights) needed to implement their recommended estimation procedures. The most complicated aspect is how to interpret or correctly use teacher-level data, such as in estimating the proportion of teachers who have recently had some form of professional development.

Generally, teachers sampled in the CNAs are not representative of the population of teachers but rather are representative of the teachers of students. For example, one could estimate the percent of students whose teacher has recently completed professional development but not the percent of teachers who have recently completed professional development. There are also some additional computational challenges with using the teacher data, but these procedures are well documented by the CNAs and discussed below.

Table 13. CNAs, target populations and methods for computing standard errors

Assessment	Method	Reference for formulas	Target population
LLECE 2013	Linearized	StataCorp 2013	Third- and sixth-grade students
PASEC 2014/19	Jackknife repeated replication	PASEC 2017	Second- and sixth-grade students
PISA	Balanced repeated replication	OECD 2009	15-year-old students
SEA-PLM 2019	Jackknife repeated replication	SEA-PLM 2020	Fifth-grade students
TIMSS/PIRLS	Jackknife repeated replication	Foy & LaRoche (2016)	Fourth- and eighth-grade students

Source: UIS

Dataset structure: Datasets provided by the CNA reflect the sampling design. Typically, there are a student dataset, which has one row for each student containing their responses to the student questionnaire; a teacher dataset, which has one row for each teacher sampled; and a school director dataset, which has one row for each school. Linking the school dataset with the student dataset is straightforward by using the school identifier in both files. More complicated is merging the teacher dataset with the student dataset because often more than one teacher is sampled for each student. For example, TIMSS surveys students' math and science teachers and provides a student-teacher link file that has one record for each combination of student and teacher along with weights to adjust for multiple records per student. These

procedures are generally well documented in the guidebooks provided by the CNA administrators.

7.2. Process for estimating an indicator

The following section outlines the steps (with examples) for how to estimate an SDG using a CNA. It assumes this is the first indicator being created with a particular CNA; as a result, for subsequent indicators being estimated with the same CNA, some of the steps are redundant. The general approach is to create a binary variable indicating whether the student (or teacher or school) satisfies the SDG indicator in question (e.g. speaks the language of instruction at home) and then estimate the mean of that binary variable to get a proportion of the population.

Step 1: Identify variable(s) related to the indicator of interest and map responses

The first step is to identify the question items in the background questionnaires that can be used to create an estimate of the particular SDG indicator. Generally, SDG indicators refer to a proportion of a population (e.g. students) exhibiting some kind of characteristic (e.g. having been bullied). Estimating the proportion requires classifying each respondent as either having the characteristic, not having the characteristic, or not having information to determine. The estimated proportion of the population uses the sample for which having the characteristic is known.

For SDG 4.a.2 (Percentage of students experiencing bullying in the last 12 months), for example, question 38 in the PISA 2018 student questionnaire (Figure 4) could be used. In this case, the question actually comprises six questions, and based on the research literature on bullying, a child experiencing any of these six qualifies as being exposed to bullying. As a result, those students who answered "never" or "almost never" to all six would be classified as not having been bullied in the past year while those who answered any other response to any of the six would be classified as having been bullied.

Note that some leeway in definition is required because "never or almost never" (presumably meaning less than a few times a year) is grouped together. Some children who have experienced some bullying would therefore be classified as not having been bullied. As for missing responses, in the case when an indicator is based on a single question, then a missing or invalid response is excluded from the estimate. In the case of this example, the bullying question from PISA 2018 (Figure 4), it is possible that some but not all questions are unanswered. In this case, the strategy used was to define a student as not having been bullied if she or he responded never or almost never to all of these six questions that she or he responded to. The student would be defined as having been bullied if she or he responded with any other valid choice to any question.

Figure 4. PISA 2018 question used for bullying

ST038 **During the past 12 months, how often have you had the following experiences in school?**

(Some experiences can also happen in social media.)

(Please select one response in each row.)

		Never or almost never	*A few times a year*	*A few times a month*	*Once a week or more*
ST038Q03NA	Other students left me out of things on purpose.	☐01	☐02	☐03	☐04
ST038Q04NA	Other students made fun of me.	☐01	☐02	☐03	☐04
ST038Q05NA	I was threatened by other students.	☐01	☐02	☐03	☐04
ST038Q06NA	Other students took away or destroyed things that belonged to me.	☐01	☐02	☐03	☐04
ST038Q07NA	I got hit or pushed around by other students.	☐01	☐02	☐03	☐04
ST038Q08NA	Other students spread nasty rumours about me.	☐01	☐02	☐03	☐04

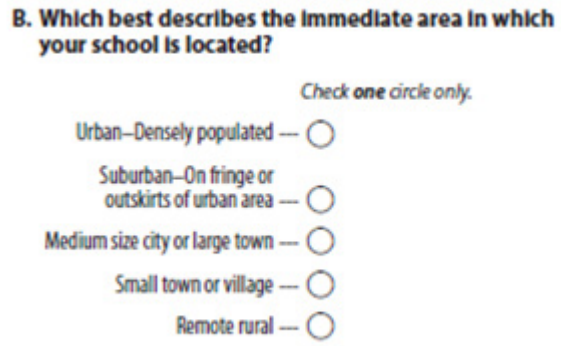
Source: PISA

Note that there are variables in the dataset that have been derived from those in the questionnaire and that might be more appropriate to use for an SDG indicator. For example, whether or not the student speaks the language of instruction (i.e. of the test) at home is often included in the dataset based on the student's responses to questions about language spoken at home. Information about this variable would be included in the codebook but not the questionnaire.

Step 2: Identify subpopulation variables

The second step is to identify the questions in the questionnaires that could be used to define subpopulations of interest. The subpopulations of interest are typically the sex of the student (or teacher), urban or rural location, and whether the student has relatively high or low SES. The sex of the student is straightforward. Defining whether the student should be classified as being urban or rural is based on the school questionnaire, which asks about the location of the school. The approach is the same as described in Step 1: identify the appropriate question and map the responses to a subpopulation or to a missing category. For example, the urban/rural question in TIMSS (Figure 5) asks the school director what best describes the area where the school is located. For this question, urban or sub-urban was defined as urban, while the others were defined as being rural.

Figure 5. Urban/rural question from TIMSS 2015



Source: TIMSS

Defining whether a student belongs to a high- or low-SES group is more involved. The approach used was to take an indicator of SES that was a continuous variable and assign those above median to being high SES and those below median to being low SES. Here some judgement is required. For example, PISA has a wealth index and an economic, social and cultural index (the ESCS), the latter of which the OECD tends to use for discussing SES differences in its reports. As a result, this latter index was used to define high and low SES. For TIMSS, no SES index exists; as a result, the index of learning materials available at home was used as this was closest to an SES measure.

Step 3: Create the dataset of interest

CNA data are usually distributed in different datasets depending on the respondent; for example, there may be one dataset for students, one or more datasets for teachers, and one dataset for schools. In some cases, the student dataset may be split into more than one dataset with item responses in one dataset and background characteristics in another. Based on Steps 1 and 2, which of these datasets are required to estimate the SDG indicator will be determined. Because the urban and rural subpopulation identifier will be defined using the school-level dataset, the school-level and student-level datasets will need to be merged. This process is straightforward and can be done using the school identifier in both datasets. In some cases, the data will be distributed in this format, as is the case for PASEC.

Working with the teacher data is more complicated when more than one teacher per student has been sampled. In these cases, a file may exist, as in TIMSS, which has one record for each student–teacher combination; for example, if two teachers of a particular student have been surveyed, then there will be two rows in the link file for that student, one for each of the teachers. This file in TIMSS includes a teacher weight (tchwgt), which is the student weight divided by the number of times the student appears in the link file. The student file is then merged into the student-teacher link file and the school-level dataset can also be merged in if needed. Estimating student characteristics using this file with the student-teacher link weight should produce the same results as estimating characteristics using the student file with the original student weight (assuming the estimation method suggested by TIMSS is applied).

In the case of LLECE 2013, it was necessary to create this student-teacher link file manually. In this assessment, there may be up to three teachers surveyed for each student. To build the link file, the student-level data was cleared except for the student code, school code and weights. Three variables were created for each of the three teachers, and then the dataset was reshaped, using Stata, to long format with three rows for each student. The three teacher datasets were merged successively based on the school code. Missing rows – that is, where there were fewer than three teachers for a student – were removed. The student-teacher link weight was created by dividing the student weight by the number of times the student appeared in the link file. Note that implementing this requires codes that are unique across countries or implemented separately for each country.

Step 4: Create the subpopulation variables

Creating the subpopulation variables is generally straightforward. Here, for example, is the Stata code for creating the urban/rural identifier in TIMSS 2015:

```
gen urban = 1 if bcbg05b != “NA”
replace urban = 0 if bcbg05b == “Remote rural” | bcbg05b == “Small town or village”
```

The SES indicator variable is more complicated because it requires identifying the median value first. The Stata code for converting the PISA ESCS variable is

```
replace escs = escs if escs < 95
sum escs [aw=w_fstuwt], d
gen highses = 0 if escs <= r(p50)
replace highses = 1 if escs > r(p50) & escs < .
```

The first line removes missing values (coded as 95 or higher). The second step is to calculate the population median, and the third and fourth steps are to create the high-SES indicator variable based on the population median.

Step 5: Create the variable for the SDG indicator

Creating the indicator variable for the SDG indicator may be complicated when the indicator depends on more than one question variable in the dataset. Consider this code used to generate the bullied indicator variable for PISA 2018:

```
qui gen bullied = .
foreach var of varlist *038* {
qui replace bullied = 1 if (`var’ == “A few times a month” | `var’ == “A few times a
year” | `var’ == “Once a week or more”)
qui replace bullied = 0 if (`var’ == “Never or almost never”) & bullied >= .
}
```

Here, a loop has been used because of the number of questions related to bullying. In this dataset, only the variables needed for the analysis were included, and as a result only the variables with 038 in their

name were the bullying variables. The code begins by setting the bullied indicator to missing and then loops through each of the bullying variables. It will replace the bullying variable with "1" if the student chose any response to the variable in the loop other than "never or almost never", and it will replace the bullied indicator with "0" if it is missing and the student selected "never or almost never".

Step 6: Estimate the proportion

The challenge in estimating the proportion is to understand how to implement the estimation method recommended by the administrators of the assessment (Table 10 provides the name of and reference for the estimation methods for each of the major CNAs). Generally, the estimation methods employed by the CNAs are designed to provide unbiased or statistically consistent estimates of the indicator and its standard error. Because of the complex survey design, most CNAs recommend a resampling method to estimate the standard errors. A standard error is an estimate of the sampling variation of the indicator and can be measured, for example, by conducting the survey repeatedly. Of course, such repetition is not feasible, and a resampling approach imitates this design by drawing subsamples of the sample and calculating the standard error from the variation in the estimated indicator for each of the subsamples. The CNAs generally include the variables needed to implement their recommended method for estimating the indicator and its standard error. PISA and PASEC, for example, includes a set of replicate weights that define each subsample, while TIMSS includes the variables needed to manually define the subsamples. Subroutines have been written for implanting these codes. The following are some examples using Stata for estimating the proportion of students who speak the language of the test at home (the definition of this variable for each of these assessments is included in the Appendix).

For PISA 2018, the Plausible Value (PV) module is used:

```
pv, pv(testlanghome) brr fays(0.5) weight(w_fstuwt) rw(w_fsturwt*): reg @pv [aw=@w]
```

Using this code, the proportion of testlanghome is estimated using the student weight, w_fstuwt, and the standard error is estimated using balanced repeated replication using the replicate weights, w_fsturwt1 - w_fsturwt80. The reg command was used to estimate the proportion; this provides equivalent results to using mean but is generally faster to execute in Stata. Note that only the point estimates from the replicate weights are used in calculating the standard errors; the standard errors produced by the reg command are not used.

For TIMSS 2019, the PV module is also used but the command is changed to match the method used by TIMSS:

```
pv, pv(testlanghome) jkzone(jkzone) jkrep(jkrep) jrrt2 weight(totwgt): reg @pv [aw=@w]
```

By this specification, the PV module estimates the testlanghome, creates the jackknife replicate weights and estimates the standard error using the estimates of testlanghome for each replicate weight.

For PASEC 2019, the jackknife replicate weights are included in the dataset; the estimates and jackknife standard errors are obtained by

```
pv, pv(testlanghome) jrr rw(rwgt1 - rwgt90) weight(rwgt0) jk(2): reg @pv [aw=@w]
```

The estimates for subpopulations can be obtained by adding an if statement to the commands, for example as

```
pv, pv(testlanghome) brr fays(0.5) weight(w_fstuwt) rw(w_fsturwt*): reg @pv [aw=@w] if
urban == 0
```

For school-level data, the estimation method is simpler because schools are the primary sampling unit and there is no intracluster correlation. Unequal school selection probabilities need to be accounted for, and school-level weights are generally included in the dataset.

Estimating teacher-level indicators is analogous to estimating student-level indicators for assessments where more than one teacher is sampled per student, as in TIMSS and LLECE 2013. The difference is in the weights used. For these, the teacher indicators are not representative of the teachers in the country but rather the teachers of the students – in other words, what proportion of students have teachers with a particular characteristic, such as having received professional development in the past 12 months. As discussed previously, the student-teacher link file is used as the basis of the dataset if there is more than one teacher per student. The weight is the student weight divided by the number of times the student appears in the data (e.g. normally twice in eighth-grade TIMSS, once linked to her or his science teacher and once to her or his math teacher). For example, to estimate the proportion of students with a teacher who has received professional development in the past 12 months, the tchwgt weight would be used, as in this example code:

```
pv, pv(recentpd) jkzone(jkzone) jkrep(jkrep) jrrt2 weight(tchwgt): reg @pv [aw=@w]
```

For PISA, the teacher sample was drawn randomly from teachers within the school who met their target population of teachers. Estimates from this sample are representative of the target population of teachers. Estimating statistics using these data requires methods that are robust for intracluster correlation (e.g. using Stata's built-in survey set commands).

7.3. Further considerations

Validating the estimation methodology: Generally it is good practice to validate that the estimation method specified in the analysis software (e.g. the options used in module PV in the examples above) is correct. One way to test this is to estimate the test scores and check whether the estimates and standard errors match those in the assessment's report. Test scores are generally reported using multiple imputations or plausible values; the idea is that a student's test score is innate and, as a result, the achievement scores estimated based on her or his answers to the test questions have uncertainty and are treated as a probability distribution. The plausible values provide a distribution, and statistics based on achievement scores (for example, the mean) have a standard error that reflects not only the sampling variation but also the variation among the plausible values resulting from uncertainty around the students' achievement scores. Instructions for calculating estimates using the plausible values are described in the various assessments' technical reports or user manuals, and subroutines for statistical software exist to implement them. For example, to estimate average math achievement for a PISA country using the PV module for Stata, the command would be

```
pv, pv(pv*math) brr fays(0.5) weight(w_fstuwt) rw(w_fsturwt*): reg @pv [aw=@w]
```

Comparing these results from a country and by subpopulation to ensure they match those of the report or statistical almanac helps ensure that the analyst has correctly specified the estimation method.

Data management: If the analyst is generating estimates for a large number of countries as well as subpopulations within these countries, then procedures for storing and managing these estimates will need to be developed. A convenient aspect of international learning assessment data is that the variables are standardized across countries, so estimates can be generated for each country and subpopulation using loops. The code would need to take into account datasets that may not have any responses to a particular indicator or subpopulation. A method for storing the output of these estimates, including the point estimate, standard error and confidence intervals, would need to be established. As part of this, decision rules would need to be determined to handle cases with few observations for a particular subpopulation and to determine whether estimates from these should be flagged.

7.4. How the metadata helps

Metadata describing how CNA data were used to generate SDG 4 estimates are essential to both understand the data and transparency in monitoring the SDGs. In order to achieve this objective of transparency,

metadata should provide sufficient information to allow researchers to replicate the reported estimates. Indicators derived from CNAs (and other surveys) are unique in that they are estimates. Technically, these estimates are random variables with probability distributions resulting from the random sampling design. Metadata for all SDG indicators could be found in the TCG microsite.

Metadata documentation should include the following points:

- The questions that were used to define an indicator and how the responses were mapped to the indicator definitions;
- The target population of the data used;
- Discussion on the comparability of the indicator with indicators derived from different assessment programmes;
- How the target populations of the various assessment programmes are mapped to the levels defined in each indicator (e.g. eighth grade to lower secondary);
- How instances of more than one target population per country per year can be mapped to a single data point per country per year (e.g. school-level indicators for primary school derived from second and sixth grade in the case of PASEC).

Appendix 1 presents the questions used to derive the SDG 4 indicators currently reported that are not measures of learning outcomes.

Finally, metadata points would ideally provide, in addition to the source of data and reference to the appropriate metadata documentation, an indication of sampling variation. Each estimated value for an SDG indicator derived from sample-based data sources has a probability distribution. One approach would be to present the 95% confidence interval (the range of values around the estimate within which there is a 95% probability of the true indicator value being) in order to specify the possibility of sampling error.

8. References

ACER/UIS. https://www.acer.org/au/gem/learning-progression-explorer. Accessed 7 January 2021.

Altinok, N (2017). Mind the Gap: Proposal for a Standardised Measure for SDG 4– Education 2030 Agenda. Montreal: UNESCO Institute for Statistics (UIS). Information Paper 46. http://uis.unesco.org/sites/default/files/documents/unesco-infopaper-sdg_data_gaps-01.pdf

Dumais, J., & Gough, H. (2012). School sampling methodology. In V. Greaney & T. Kellaghan (Eds.), Implementing a National Assessment of Educational Achievement. Chichester, UK: World Bank.

Foy, P. and S. LaRoche (2016). Chapter 4: Estimating standard errors in the TIMSS 2015 results. M.O. Martin, I.V.S. Mullis and M. Hooper (eds), Methods and Procedures in TIMSS 2015. Boston, Massachusetts, USA: Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-4.html

Greaney, V and T. Kellaghan (2008). Assessing National Achievement Levels in Education. National Assessments of Educational Achievement. Washington, DC: World Bank. World Bank. https://openknowledge.worldbank.org/handle/10986/6904 License: CC BY 3.0 IGO.

OECD (2009). PISA Data Analysis Manual SPSS, Second Edition. Paris: OECD. https://www.oecd-ilibrary.org/docserver/9789264056275-enpdf?expires=1593772556&id=id&accname=guest&checksum=256201EAFD784D9168955CFDB6068789

PASEC (2017). Manuel d'exploitation des données : Évaluation internationale PASEC2014. Dakar: PASEC, CONFEMEN. https://www.pasec.confemen.org/wp-content/uploads/2017/06/Manuel_exploitation_donn%C3%A9es_%C3%A9valuation_internationale_PASEC2014.pdf

Rocher, Tand D. Hastedt (2020). International Large-Scale Assessments in Education: A Brief Guide. In: Compass Briefs in Education. International Association for the Evaluation of Educational Achievement (IEA) – Number 10 – September 2020. https://www.iea.nl/sites/default/files/2021-07/2021.07.20_ILSAs%20in%20education%20-%20a%20brief%20guide%20Compass%2010.pdf

Rosetta Stone (at UIS website). http://tcg.uis.unesco.org/rosetta-stone/. Accessed 7 January 2021.

Rust, K. F. (2014). Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of International Large-Scale Assessment Background. Technical Issues, and Methods of Data Analysis. Boca Raton, London, New York: CRC Press.

Rust, K. F., Krawchuk, S., & Monseur, C. (2017). Sample Design, Weighting, and Calculation of Sampling Variance. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), Implementation of Large-Scale Education Assessments. Chichester, UK: Wiley.

SEA-PLM (2020). Data User Manual. Bangkok: SEAMEO and UNICEF.

StataCorp (2013). Stata Survey Data Reference Manual Release 13. College Station, Texas, USA: Stata Press. https://www.stata.com/manuals13/svy.pdf

UNESCO Institute for Statistics (2016). Sustainable Development Data Digest: Laying the Foundation to Measure Sustainable Development Goal 4. Montreal: UNESCO Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/laying-the-foundation-to-measure-sdg4-sustainable-development-data-digest-2016-en.pdf

United Nations (2014). The road to dignity by 2030: ending poverty, transforming all lives and protecting the planet. Synthesis report of the Secretary-General on the post-2015 sustainable development agenda. https://www.un.org/ga/search/view_doc.asp?symbol=A/69/700&Lang=E

Appendix 1: Mapping of questionnaire responses used for defining SDG 4 indicators

Table 14. Questionnaires and mapping of responses for SDG 4.5.2 (percent of students learning in their home language)

Data source	Target population (measurement point)	Language at home questions	Responses and mapping to whether the student uses the language of the test at home (yes/no/ omitted)
LLECE 2013 (TERCE)	6th-grade students (end of primary)	At home, which language do you speak most of the time?	"Spanish or Portuguese": yes All other valid responses: no Missing: omitted
PASEC 2014	2nd-grade students (Grades 2/3)	What language do you speak at home?	"You always speak <language of test>": yes "You speak <language of test> sometimes and another language sometimes": no "You never speak <language of test>": no Missing: omitted
PASEC 2014/2019	6th-grade students (end of primary)	What language do you speak at home?	"You always speak <language of test>": yes "You speak <language of test> sometimes and another language sometimes": no "You never speak <language of test>": no Missing: omitted
PISA 2018	15-year-old secondary students (end of lower second-ary)	What language do you speak at home most of the time? (please se-lect one response)	"<Language 1>" "<Language 2>" "<Language 3>" "<…etc.>" "Other languages" Assignment of these responses to whether the student speaks the language of the test at home most of the time is done by the OECD and re-ported as a variable in the dataset.
SEA-PLM 2019	5th-grade students (end of primary)	What language do you speak at home most of the time? Note if two languages are spoken at the same frequency, choose the one you learnt first.	"<Language 1>" "<Language 2>" "<Language 3>" "<Language 4>" "<Other language>" Assignment of these responses to whether the student speaks the language of the test at home most of the time is done by SEA-PLM and re-ported as a variable in the dataset.
TIMSS 2015/2019	4th-grade students (Grades 2/3)	How often do you speak <language of test> at home? Fill one circle on-ly.	"I always speak <language of test> at home": yes "I almost always speak <language of test> at home": yes I sometimes speak <language of test> and sometimes speak another language at home": no "I never speak <language of test> at home": no Missing or invalid: omitted
TIMSS 2015/2019	8th-grade students(end of lower second-ary)	How often do you speak <language of test> at home? Fill one circle on-ly.	"Always": yes "Almost always": yes "Sometimes": no "Never": no Missing or invalid: omitted

Source: UIS

Appendix 2: Stata codes used to define SDG indicator variables

(for SDG indicators 4.5.2, 4.a.1, 4.a.2 and 4.c.7 in the UIS database as of September 2021)

Note: indicators are only listed for an assessment if it is possible to calculate the indicator; otherwise, the indicator is omitted. In addition, they are generally specified using datasets that have removed unneeded variables and for only one country. Each major assessment programme is indicated in bold.

LLECE 2013

3rd Grade: 4.a.1 (electricity)

```
gen electricity = dqdit17_01 if dqdit17_01 == 0 | dqdit17_01 == 1
```

3rd Grade: 4.a.1 (drinking water)

```
gen water = dqdit17_02 if dqdit17_02 == 0 | dqdit17_02 == 1
```

3rd Grade: 4.a.1 (computers for pedagogic purposes)

```
gen computer = 0 if dqdit15_01 <= 5
replace computer = 1 if dqdit15_01 >= 2 & dqdit15_01 <= 5
replace computer = 1 if dqdit15_02 >= 2 & dqdit15_02 <= 5
```

3rd Grade: 4.a.1 (internet for pedagogic purposes)

```
gen internet = 0 if dqdit15_02 <= 5
replace internet = 1 if dqdit15_02 >= 2 & dqdit15_02 <= 5
```

3rd Grade: SDG 4.c.7 (professional development in the past 12 months)

```
gen recentpd = 1 - dqpit12_07 if dqpit12_07 == 0 | dqpit12_07 == 1
```

3rd Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if dqpit02 == 1
replace female_teacher = 1 if dqpit02 == 2
```

6th Grade: SDG 4.5.2 (language of instruction used at home)

```
gen testlanghome = 0 if dqa6it05 <= 6
replace testlanghome = 1 if dqa6it05 == 1
```

6th Grade: SDG 4.a.1 (electricity)

```
gen electricity = dqdit17_01 if dqdit17_01 == 0 | dqdit17_01 == 1
```

6th Grade: SDG 4.a.1 (drinking water)

```
gen water = dqdit17_02 if dqdit17_02 == 0 | dqdit17_02 == 1
```

6th Grade: SDG 4.a.1 (computer for pedagogic use)

```
gen computer = 0 if dqdit15_01 <= 5
replace computer = 1 if dqdit15_01 >= 2 & dqdit15_01 <= 5
replace computer = 1 if dqdit15_02 >= 2 & dqdit15_02 <= 5
```

6th Grade: SDG 4.a.1 (internet)

```
gen internet = 0 if dqdit15_02 <= 5
replace internet = 1 if dqdit15_02 >= 2 & dqdit15_02 <= 5
```

6th Grade: SDG 4.a.2 (bullying)

```
qui gen bullied = .
foreach var of varlist dqa6it19_0? {
qui replace bullied = 0 if `var' == 0 & bullied >= .
qui replace bullied = 1 if `var' == 1
}
```

6th Grade: SDG 4.c.7 (teachers' recent professional development)

```
gen recentpd = 1 - dqpit12_07 if dqpit12_07 == 0 | dqpit12_07 == 1
```

6th Grade: subpopulation (female/male)

```
gen female = 0 if dqa6it02 == 2
replace female = 1 if dqa6it02 == 1
```

6th Grade: subpopulation (urban/rural)

```
gen urban = 1 if dqdit12 == 2
replace urban = 0 if dqdit12 == 1
```

6th Grade: subpopulation (low/high SES)

```
replace isecf = . if isecf == 9
sum isecf [aw=wgt], d
gen highses = 0 if isecf <= r(p50)
replace highses = 1 if isecf > r(p50) & isecf < .
```

6th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if dqpit02 == 1
replace female_teacher = 1 if dqpit02 == 2
```

PASEC 2014

2nd Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if qe25 == 2 | qe25 == 3
replace testlanghome = 1 if qe25 == 1
```

2nd Grade: SDG 4.a.1 (electricity)

```
gen electricity = 2 - qd65o if qd65o >= 1 & qd65o <= 2
```

2nd Grade: SDG 4.a.1 (drinking water)

```
gen water = 2 - qd65m if qd65m >= 1 & qd65m <= 2
replace water = 2 - qd65l if (water >= . | water == 0) & qd65l >= 1 & qd65l <= 2
```

2nd Grade: SDG 4.c.7 (teachers' recent professional development)

```
gen recentpd = 2 - qm27a if qm27a >= 1 & qm27a <= 2
```

2nd Grade: subpopulation (female/male)

```
gen female = 0 if qe22 == 1
replace female = 1 if qe22 == 2
```

2nd Grade: subpopulation (urban/rural)

```
gen urban = 1 if qd24 == 1 | qd24 == 2
replace urban = 0 if qd24 == 3 | qd24 == 4
```

2nd Grade: sub-population – female/male teacher

```
gen female_teacher = 0 if qm21 == 1
replace female_teacher = 1 if qm21 == 2
```

6th Grade: SDG 4.5.2 (language of instruction used at home)

```
gen testlanghome = 0 if qe615 == 2 | qe615 == 3
replace testlanghome = 1 if qe615 == 1
```

6th Grade: SDG 4.a.1 (electricity)

```
gen electricity = 2 - qd65o if qd65o >= 1 & qd65o<= 2
```

6th Grade: SDG 4.a.1 (drinking water)

```
gen water = 2 - qd65m if qd65m >= 1 & qd65m <= 2
replace water = 2 - qd65l if (water >= . | water == 0) & qd65l >= 1 & qd65l <= 2
```

6th Grade: SDG 4.c.7 (teachers' recent professional development)

```
gen recentpd = 2 - qm67a if qm67a >= 1 & qm67a <= 2
```

6th Grade: subpopulation (female/male student)

```
gen female = 0 if qe62 == 1
replace female = 1 if qe62 == 2
```

6th Grade: subpopulation (urban/rural)

```
gen urban = 1 if qd24 == 1 | qd24 == 2
replace urban = 0 if qd24 == 3 | qd24 == 4
```

6th Grade: subpopulation (low/high socio-economic status [SES])

```
sum ses [aw=rwgt0], d
gen highses = 0 if ses <= r(p50)
replace highses = 1 if ses > r(p50) & ses < .
```

6th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if qm61 == 1
replace female_teacher = 1 if qm61 == 2
```

PASEC 2019

2nd Grade: SDG 4.a.1 (electricity)

```
gen electricity = 2 - qd78o if qd78o >= 1 & qd78o <= 2
```

2nd Grade: SDG 4.a.1 (drinking water)

```
gen water = 2 - qd78m if qd78m >= 1 & qd78m <= 2
replace water = 2 - qd78n if (water >= . | water == 0) & qd78n >= 1 & qd78n <= 2
```

6th Grade: SDG 4.a.1 (electricity)

```
gen electricity = 2 - qd78o if qd78o >= 1 & qd78o <= 2
```

6th Grade: SDG 4.a.1 (drinking water)

```
gen water = 2 - qd78m if qd78m >= 1 & qd78m <= 2
replace water = 2 - qd78n if (water >= . | water == 0) & qd78n >= 1 & qd78n <= 2
```

6th Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if qe620 >= 1 & qe620 <= 4
replace testlanghome = 1 if qe620 == 1
```

6th Grade: subpopulation (female/male student)

```
gen female = 0 if qe63 == 1
replace female = 1 if qe63 == 2
```

6th Grade: subpopulation (urban/rural)

```
gen urban = 1 if qd31 == 1 | qd31 == 2
replace urban = 0 if qd31 == 3 | qd31 == 4
```

6th Grade: subpopulation (high/low SES)

```
sum ses [aw=rwgt0], d
gen highses = 0 if ses <= r(p50)
replace highses = 1 if ses > r(p50) & ses < .
```

PISA 2015

SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = .
replace testlanghome = 0 if st022q01 == 2
replace testlanghome = 1 if st022q01 == 1
```

SDG 4.a.1 (computer)

```
gen computer = 0 if sc004q02ta < 99995
replace computer = 1 if sc004q02ta > 0 & computer == 0
```

SDG 4.a.1 (internet)

```
gen internet = 0 if sc004q03ta < 99995
replace internet = 1 if sc004q03ta > 0 & internet == 0
```

4.c.7 (teachers' recent professional development)

```
gen recentpd = .
foreach v of varlist tc020* {
```

```
replace recentpd = 0 if `v' == 2 & recentpd >= .
replace recentpd = 1 if `v' == 1
}
```

subpopulation (female student)

```
gen female = 0 if st004d01 == 2
replace female = 1 if st004d01 == 1
```

subpopulation (urban/rural)

```
gen urban = 1 if sc001q01ta == 3 | sc001q01ta == 4 | sc001q01ta == 5
replace urban = 0 if sc001q01ta == 1 | sc001q01ta == 2
```

subpopulation (high/low SES)

```
replace escs = escs if escs < 95
sum escs [aw=w_fstuwt], d
gen highses = 0 if escs <= r(p50)
replace highses = 1 if escs > r(p50) & escs < .
```

subpopulation female/male teacher)

```
gen female = 0 if tc001q01na == 2
replace female = 1 if tc001q01na == 1
```

PISA 2018

SDG 4.5.2 (language of instruction used at home)

```
gen testlanghome = .
replace testlanghome = 0 if st022q01ta == "Other language"
replace testlanghome = 1 if st022q01ta == "Language of the test"
```

SDG 4.a.2 (bullying)

```
gen bullied = .
foreach var of varlist *038* {
replace bullied = 1 if (`var' == "A few times a month" | `var' == "A few times a year" |
`var' == "Once a week or more")
replace bullied = 0 if (`var' == "Never or almost never") & bullied >= .    // only
overwrite NA, not 1
}
```

SDG 4.a.1 (computer)

```
gen computer = 0 if sc004q02ta < .
replace computer = 1 if sc004q02ta > 0 & computer == 0
```

SDG 4.a.1 (internet)

```
gen internet = 0 if sc004q03ta < .
replace internet = 1 if sc004q03ta > 0 & internet == 0
```

SDG 4.c.7 (teachers' recent professional development)

```
gen recentpd = .
foreach v of varlist *193* tc020q01na tc020q02na tc020q03na tc020q04na tc020q05na {
replace recentpd = 0 if `v' == "No" & recentpd >= .
```

```
replace recentpd = 1 if `v' == "Yes"
}
```

subpopulation (female/male student)

```
gen female = 0 if st004d01t == "Male"
replace female = 1 if st004d01t == "Female"
```

subpopulation (urban/rural)

```
gen urban = 1 if strpos(sc001q01ta,"A city (100 000 to about 1 000 000 peop") > 0 |
strpos(sc001q01ta,"A large city (with over 1 000 000 peopl") > 0 | strpos(sc001q01ta,"A
town (15 000 to about 100 000 people)") > 0
urban = 0 if strpos(sc001q01ta,"A small town (3 000 to about 15 000 peo") > 0 |
strpos(sc001q01ta,"A village, hamlet or rural area (fewer") > 0
```

subpopulation (high/low SES)

```
sum escs [aw=w_fstuwt], d
gen highses = 0 if escs <= r(p50)
replace highses = 1 if escs > r(p50) & escs < .
```

ubpopulation (female/male teacher)

```
gen female = 0 if tc001q01na == "Male"
replace female = 1 if tc001q01na == "Female"
```

SEAPLM 2019

SEAPLM 2019: SDG 4.5.2 (language of instruction spoken at home)

```
recode s_lang (1=0) (2=1) (else=.), gen(testlanghome)
```

SEAPLM 2019: SDG 4.a.1 (electricity)

```
recode sc15q05 (1=1) (2=0) (else=.), gen(electricity)
```

SEAPLM 2019: SDG 4a.1 (drinking water)

```
recode sc15q07 (1=1) (2=0) (else=.), gen(water)
```

SEAPLM 2019: SDG 4.a.1 (separate toilets)

```
recode sc20q02 (1=1) (2=0) (else=.), gen(toilet_boys)
recode sc20q03 (1=1) (2=0) (else=.), gen(toilet_girls)
gen toilet_seperate = 0 if toilet_boys == 0 & toilet_girls == 0
replace toilet_seperate = 1 if toilet_boys == 1 | toilet_girls == 1
```

SEAPLM 2019: subpopulation (female/male student)

```
recode s_gender (1=0) (2=1) (else=.), gen(female)
```

SEAPLM 2019: subpopulation (urban/rural)

```
recode sc09q01 (4/5=1) (1/3=0) (else=.), gen(urban)
```

SEAPLM 2019: subpopulation (high/low SES)

```
sum ses [aw=wt2019], d
```

```
gen highses = 0 if ses <= r(p50)
replace highses = 1 if ses > r(p50) & ses < .
```

TIMSS 2015

4th Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if asbg03 == “Never” | asbg03 == “Sometimes”
replace testlanghome = 1 if asbg03 == “Almost always” | asbg03 == “Always”
```

4th Grade: SDG 4.a.1 (computer)

```
gen computer = 0 if atbm05a == “No” & atbs04a == “No”
replace computer = 1 if atbm05a == “Yes” | atbs04a == “Yes”
collapse (first) schwgt  country (max) computer, by( idcntry idschool)
```

4th Grade: SDG 4.a.2 (bullying)

```
gen bullied = .
foreach var of varlist *sbg12* {
local a1 = “A few times a year”
local a2 = “At least once a week”
local a3 = “Once or twice a month”
local a4 = “Never”
qui replace bullied = 1 if (`var’ == “A few times a year” | `var’ == “At least once a
week” | `var’ == “Once or twice a month”)
qui replace bullied = 0 if (`var’ == “Never”) & bullied >= .
}
```

4th Grade: SDG 4.c.7 (teachers’ recent professional development)

```
gen recentpd = 1
replace recentpd = 0 if (atbm10 == “NA” & atbs09 == “None”) | (atbm10 == “None” & atbs09
== “NA”) | (atbm10 == “None” & atbs09 == “None”)
replace recentpd = . if atbm10 == “NA” & atbs09 == “NA”
```

4th Grade: subpopulation (female/male student)

```
gen female = 0 if itsex == “Male”
replace female = 1 if itsex == “Female”
```

4th Grade: subpopulation (urban/rural)

```
gen urban = 1 if acbg05b != “NA”
replace urban = 0 if acbg05b == “Remote rural” | acbg05b == “Small town or village”
```

4th Grade: subpopulation (high/low SES)

```
sum asbghrl [aw=matwgt], d
gen highses = 0 if asbghrl <= r(p50)
replace highses = 1 if asbghrl > r(p50) & asbghrl < .
```

4th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if atbg02 == “Male”
replace female_teacher = 1 if atbg02 == “Female”
```

8th Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if bsbg03 == “Never” | bsbg03 == “Sometimes”
replace testlanghome = 1 if bsbg03 == “Almost always” | bsbg03 == “Always”
```

8th Grade: SDG 4.a.1 (computer)

```
gen computer = 0 if btbs19a == “No”
replace computer = 1 if btbs19a == “Yes”
replace computer = 0 if btbm20a == “No”
replace computer = 1 if btbm20a == “Yes”
collapse (first) schwgt  country (max) computer, by( idcntry idschool)
```

8th Grade: SDG 4.a.2 (bullying)

```
gen bullied = .
foreach var of varlist *sbg16* {
local a1 = “A few times a year”
local a2 = “At least once a week”
local a3 = “Once or twice a month”
local a4 = “Never”
replace bullied = 1 if (`var’ == “A few times a year” | `var’ == “At least once a week”
| `var’ == “Once or twice a month”)
replace bullied = 0 if (`var’ == “Never”) & bullied >= .      // only overwrite NA, not
1
}
```

8th Grade: SDG 4.c.7 (teachers’ professional development)

```
gen pd = btbm25
replace pd = btbs24 if pd == “”
gen recentpd = 1
replace recentpd = 0 if pd == “None”
replace recentpd = . if pd == “” | pd == “NA”
```

8th Grade: subpopulation (female/male student)

```
gen female = 0 if itsex == “Male”
replace female = 1 if itsex == “Female”
```

8th Grade: subpopulation (urban/rural)

```
gen urban = 1 if bcbg05b != “NA”
replace urban = 0 if bcbg05b == “Remote rural” | bcbg05b == “Small town or village”
```

8th Grade: subpopulation (high/low SES)

```
sum bsbgher [aw=matwgt], d
gen highses = 0 if bsbgher <= r(p50)
replace highses = 1 if bsbgher > r(p50) & bsbgher < .
```

8th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if btbg02 == “Male”
replace female_teacher = 1 if btbg02 == “Female”
```

TIMSS 2019

4th Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if asbg03 == “I never speak <language of test> at home” | asbg03 ==
```

```
“I sometimes speak <language of test> and sometimes speak another language at home”
replace testlanghome = 1 if asbg03 == “I almost always speak <language of test> at home”
| asbg03 == “I always speak <language of test> at home”
```

4th Grade: SDG 4.a.1 (computer)

```
gen computer = 0 if atbm04a == “No” & atbs03a == “No”
replace computer = 1 if atbm04a == “Yes” | atbs03a == “Yes”
collapse (first) schwgt  country (max) computer, by( idcntry idschool)
```

4th Grade: SDG 4.a.2 (bullying)

```
qui gen bullied = .
foreach var of varlist asbg11* {
local a1 = “A few times a year”
local a2 = “At least once a week”
local a3 = “Once or twice a month”
local a4 = “Never”
qui replace bullied = 1 if (`var’ == “A few times a year” | `var’ == “At least once a
week” | `var’ == “Once or twice a month”)
qui replace bullied = 0 if (`var’ == “Never”) & bullied >= .
}
```

4th Grade: SDG: 4.c.7 (teachers’ professional development)

```
gen recentpd = 1
replace recentpd = 0 if (atbm10 == “NA” & atbs09 == “None”) | (atbm10 == “None” & atbs09
== “NA”) | (atbm10 == “None” & atbs09 == “None”)
replace recentpd = . if atbm10 == “NA” & atbs09 == “NA”
```

4th Grade: subpopulation (female/male student)

```
gen female = 0 if itsex == “Male”
replace female = 1 if itsex == “Female”
```

4th Grade: subpopulation (urban/rural)

```
gen urban = 1 if acbg05b != “NA”
replace urban = 0 if acbg05b == “Remote rural” | acbg05b == “Small town or village”
```

4th Grade: subpopulation (high/low SES)

```
sum asbghrl [aw=`w’], d
gen highses = 0 if asbghrl <= r(p50)
replace highses = 1 if asbghrl > r(p50) & asbghrl < .
```

4th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if atbg02 == “Male”
replace female_teacher = 1 if atbg02 == “Female”
```

8th Grade: SDG 4.5.2 (language of instruction spoken at home)

```
gen testlanghome = 0 if bsbg03 == “Never” | bsbg03 == “Sometimes”
replace testlanghome = 1 if bsbg03 == “Almost always” | bsbg03 == “Always”
```

8th Grade: SDG 4.a.1 (computer)

```
gen computer = 0 if btbs16a == “No”
replace computer = 1 if btbs16a == “Yes”
replace computer = 0 if btbm17a == “No”
```

```
replace computer = 1 if btbm17a == “Yes”
collapse (first) schwgt  country (max) computer, by(idcntry idschool)
```

8th Grade: SDG 4.a.2 (bullying)

```
gen bullied = .
foreach var of varlist bsbg14* {
local a1 = “A few times a year”
local a2 = “At least once a week”
local a3 = “Once or twice a month”
local a4 = “Never”
qui replace bullied = 1 if (`var’ == “A few times a year” | `var’ == “At least once a
week” | `var’ == “Once or twice a month”)
qui replace bullied = 0 if (`var’ == “Never”) & bullied >= .
}
```

8th Grade: SDG: 4.c.7 (teachers’ professional development)

```
gen recentpd = 1
replace recentpd = 0 if btbm23 == “None”
replace recentpd = . if btbm23 == “NA”
replace recentpd = 0 if btbs22 == “None”
replace recentpd = . if btbs22 == “NA”
```

8th Grade: subpopulation (female/male student)

```
gen female = 0 if itsex == “Male”
replace female = 1 if itsex == “Female”
```

8th Grade: subpopulation (urban/rural)

```
gen urban = 1 if bcbg05b != “NA”
replace urban = 0 if bcbg05b == “Remote rural” | bcbg05b == “Small town or village”
```

8th Grade: subpopulation (high/low SES)

```
sum bsbgher [aw=`w’], d
gen highses = 0 if bsbgher <= r(p50)
replace highses = 1 if bsbgher > r(p50) & bsbgher < .
```

8th Grade: subpopulation (female/male teacher)

```
gen female_teacher = 0 if btbg02_m == “Male” | btbg02_s == “Male”
replace female_teacher = 1 if btbg02_m == “Female” | btbg02_s == “Female”
```

Appendix 3: Resources for calculation - Stata codes used to estimate SDG indicator variables

This section provides the codes used to estimate the indicators based on their definitions above. Examples are given for each type of data source (student-level, teacher-level, and school-level) as relevant. Data is assumed to be restricted to a specific country or sub-population within the country. One example is given per assessment program unless the estimation method differs between assessment years.

LLECE 2013

Student-level data (one record per student, dataset is restricted to a single country)

```
egen cluster = group(idcntry idschool)
svyset cluster [pw=wgt]
svy: reg testlanghome
```

Teacher-level data (one record per combination of student and teacher with each student weight divided by the number of teachers per student).

```
svyset cluster [pw=wgt]
svy: reg recentpd
```

School-level data (one record per school)

```
reg water [aw=bsw]
```

PASEC

Student-level data (one record per student, restricted to a single country)

```
pv, pv(testlanghome) jrr rw(rwgt1 - rwgt90) weight(rwgt0) jk(2): reg @pv [aw=@w]
```

School-level data (one record per school, restricted to a single country)

```
reg electricity [aw= w_sch_adj6]
```

Note: weight variable varies by year of assessment and grade level.

PISA

Student-level data (one record per student)

```
pv, pv(bullied) brr fays(0.5) weight(w_fstuwt) rw(w_fsturwt*): reg @pv [aw=@w]
```

Teacher level data (one record per teacher)

```
svyset cntschid [pw=w_schgrnrabwt]
svy: reg recentpd
```

School-level data

```
svyset [pw=w_schgrnrabwt]
svy: reg internet
```

SEA-PLM

Student-level data (one record per student)

```
pv, pv(testlanghome) jrr jk(2) rw(rwgt*) weight(wt2019): reg @pv [aw=@w]
```

Note: school-level indicators were estimated using student-level data (see 4.a.1 metadata).

TIMSS

Student-level data (one record per student)

```
pv, pv(bullied) jkzone(jkzone) jkrep(jkrep) jrrt2 weight(totwgt): reg @pv [aw=@w]
```

Teacher-level data (one record per teacher per student)

```
pv, pv(recentpd) jkzone(jkzone) jkrep(jkrep) jrrt2 weight(tchwgt): reg @pv [aw=@w]
```

School-level data (one record per school)

```
reg computer [aw=schwgt]
```

Institute for Statistics