Implementation Strategy for measuring SDG4 Learning Outcomes Targets

Concept Note

May 4th, 2016

Table of Contents

1. Rationale and objectives

Through Sustainable Development Goal (SDG) 4, the international community has adopted a life cycle approach to education and skills. The Education 2030 agenda pledges to "ensure inclusive and equitable quality education and promote lifelong learning for all." Five of the seven education targets in SDG 4 focus on learning outcomes for children, young people and adults (Box 1). Now, with the recent agreement on goals and a small set of global indicators, attention is turning towards implementation. Because of its mandate and unique position as the lead agency for education monitoring, the UIS will play a critical leadership role in mapping out measurement strategies to monitor the new education agenda.

The mandate outlined by Education 2030 will reshape global education, including the potential to significantly improve the efficiency and reliably of learning measurement. The new agenda requires efficient and accurate systems to measure multiple forms of learning across learners in and out of formal schooling; generate actionable evidence and respond to unique national and regional contexts; and inform global monitoring. Working across such diverse goals requires considerable investment, but will also provide multiple benefits.

To help define the new measurement agenda, the UIS is proposing a learning outcomes (LO) monitoring strategy. The LO monitoring strategy revolves around three main pillars: first, **conceptualizing and building a universal learning scale** to facilitate global measurement of learning; second, **establishing a data quality framework** to help guide the development and implementation of robust, reliable assessments; and third, **creating a platform for convening key stakeholders and thought leaders to provide input and guidance** throughout the process.

This strategy has been developed in response to emerging priorities for measurement of learning that arose during the development of Education 2030. Cross nationally-comparable data on education is central to achieving the vision behind Education 2030, because they provide information on progress towards goals, track equity, and create a common language for defining and discussing competency levels in learning. National and regional data are also extremely important in measuring progress towards learning, and perhaps are more directly relevant to the immediate policy choices needed to achieve Education 2030 goals. Effective strategies for global education measurement therefore must take national, regional and global data into account, and also must create approaches that measure diverse elements of learning across all ages.

This vision and commitment to a new measurement agenda are shared by many stakeholders. Significant expertise in measuring learning exists across national, regional and global entities. But at present, measurement of learning is largely uncoordinated, with data on some areas of learning, such as math and literacy, collected through a

patchwork of disconnected national, regional and global assessments, while other areas of learning not reliably measured. There are few opportunities for sharing expertise and transferring knowledge across focus areas or regions. With greater coordination and a strategic approach, more accurate and efficient measurement of learning is possible. Both technical solutions and partnerships are needed: technical solutions to envision how data from multiple sources can fit together and where new assessments are needed, and strong partnerships within the education measurement community to gather inputs from stakeholders and communicate the value of coordinated, comparable data. This note details the UIS strategy to measure the learning outcomes targets of SDG 4 as a way forward for effective action among the global education community.

Box 1. Five of the seven targets of SDG 4 focus on learning outcomes

SDG 4

Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

SDG 4 targets related to learning outcomes

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

4.2
By 2030, ensure that all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education

4.3
By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university

4.4
By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship

4.5
By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations

4.6
By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy

4.7
By 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development

2. Key measurement challenges: What is needed now and how to address the issues?

If the education community needs reliable and actionable data on learning as a critical lever to achieve Education 2030, there are **two key sets of challenges** to address: **reconciling the multiplicity of learning assessments** within a common framework and

reducing the high transaction costs currently associated with the lack of coordination and duplication of efforts.

The first set of challenges recognizes that many measurement exercises (national, regional and global) frequently collect data on basic competencies in mathematics and literacy. Yet these assessments cannot be used in an integrated manner to provide a global picture of learning. To use these national and regional data to inform global monitoring, shared technical standards and methodologies must be developed to ensure that the data are reliable. The technical solution to the global measurement problem does not require creating a universal test (which could be seen as the last resort) but the development of **a scale that links the various assessments together based on an agreed-upon set of shared quality criteria**.

The second set of challenges is linked to the political economy of learning assessments. Institutional, political and economic contexts at the system level create a multiplicity of issues that need to be addressed in a comprehensive, participatory and politically sensitive way. Any technical solution must take into account multiple viewpoints, identifying both globally-relevant areas of learning; conceptualizing how national and regional data can help inform global education measurement; striking an appropriate balance between global competences and the role of local influences and goals on education. Beyond the creation of a scale and a quality assessment framework, adoption of these global tools, as sound as they may be, will be difficult for many countries due to lack of resources and capacity, and/or lack of commitment. Under constrained resources, coordination among international actors becomes critical to prevent duplication of efforts, appropriately channel resources and support capacity building in countries so as to generate demand for quality data on their side. **These problems related to collective action need a solution to reduce transaction costs and improve coordination among actors.**

3. UIS strategy solutions to better measure learning outcomes

The UIS is proposing a series of steps to address the challenges in learning measurement to support the achievement of SDG 4. These steps include the design of a common framework to better define and align the measurement of learning, which will result in measurement scales that will coalesce national and regional data of learning and translate the information into common reporting metrics. In addition, a data quality assessment process must be put in place to ensure the robustness and homogeneity of the data collected and disseminated. These steps are summarized here and defined in greater detail in the latter sections:

- **Definition of a Universal Learning Scale (ULS)**. Under UIS' leadership, common frameworks for learning will be defined. Once agreed, these common steps will then be integrated into a universal learning reporting scale, which will serve as the backbone for supporting reliable national assessments and informing global

monitoring by linking national results to a global scale. A reporting metric will be developed to summarize learning for easy communication by policymakers and stakeholders. A key question to resolve among stakeholders is what can be defined globally through measurement, and what should be left open to local decision-making.

- **Defining a Data Quality Assurance Framework (DQAF).** For the ULS to have maximum value in national, regional and global tracking, the data collected on learning must meet statistical standards and best practices for technical rigor, comprehensiveness and representativeness. The UIS will help set technical standards for data from learning assessments. These will ensure the rigor and representativeness of analysis using the ULS, and will also help support countries in defining the characteristics of reliable national assessment systems.

- **Creation of a platform for input.** The ULS and DQAF represent new ways of thinking about measurement of learning. With any innovative idea, input from multiple stakeholders is required. The UIS will convene and support ongoing dialogue among stakeholders to help identify issues and solutions and to gain support from the wider education community for these new approaches.

3.1. The Universal Learning Scale

As a centrepiece of its strategy, the UIS proposes to build a universal learning scale, beginning with the creation of a common framework. A framework provides a common basis when elaborating a subject/learning domain, assessment, and measurement mechanism. For global usage, the framework must be comprehensive, transparent and coherent[1]:

- **Comprehensive:** The global framework should attempt to specify a full range of knowledge, skills and use to the greatest extent possible. The global framework should articulate progressive learning dimensions in which a series of reference points (or levels) can be established and the proficiency of learning can be described.

- **Transparent:** Information regarding the structure of the framework must be clearly formulated and explicit, available and readily comprehensible to users.

- **Coherent:** The description of the framework is logical and easy for users to understand.

When fully developed, this scale would outline the progression of learning competencies that would then serve as an internationally-agreed, common reference point for measurement of learning, which can be used at the discretion of countries. The

---

[1] Some principles of the common framework used here were adapted from the "Common European Framework of Reference for Languages: Learning, Teaching, Assessment", Language Policy Unit, Strasbourg. The article is available at: www.coe.int/lang-CEFR

ULS is intended to be used as point of reference for national assessments, which would facilitate a common understanding of key steps in learning while countries rely on their own standards and methodologies to build national assessments. The primary goal of this scale is to promote reliability and comparability of national, regional and cross-national assessments, while simultaneously helping to inform global monitoring.

It may be possible to measure reading and mathematics using a common scale and reporting metric sometime in the near future. Yet given the extent of variation across the targets, progression towards the ULS will be different for each target. For example, with Target 4.1 the global education community has identified three measurement points to assess the learning of children: early grades (grades 2/3), end of primary and end of lower secondary.

As a priority, universal learning scales will develop in two key areas – reading/literacy and mathematics/numeracy – because they are foundational to learning across all academic areas and relevant data are widely available. Reasonable time frames to develop scales for other targets will be assessed as the strategy develops.

Figure 1 describes a potential learning scale where students who learn at different paces would be placed at different levels in the scale. At the same time, countries that teach the basics of reading at different ages will have students at different levels on the scale.

Figure 1: An example of a learning scale
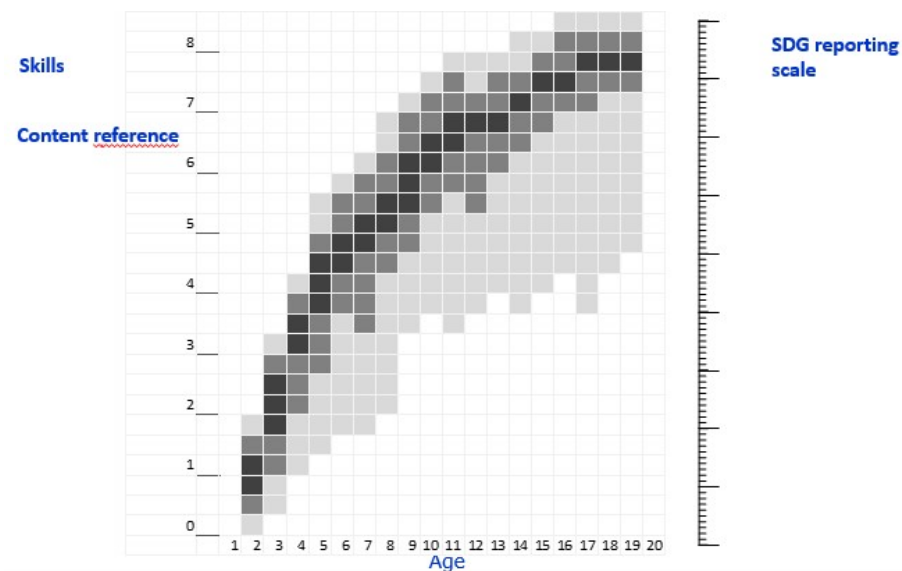
3.2. Data Quality Assurance Framework

The second pillar of the UIS strategy is to create a data quality assurance framework for learning assessments. In developing this strategy, the UIS is placing strong emphasis on the importance and value of national and regional assessment

systems. The shift in perspective from global to regional and national monitoring requires assurance that assessments systems share a basic level of quality and are developed and implemented using a common methodological framework.

The DQAF is required for two major reasons:

- **Uneven technical standards**. While many countries and regions have made tremendous progress in building and implementing learning assessments, their quality and scope vary widely due to lack of investment in measurement infrastructures, unclear standards to define equitable measurement systems, and other issues. Global standards can help support the development and implementation of reliable systems, which are critical to ensure that a wide range of learning is measured among all people in the population, not just those in school.

- **Increased coherence of data.** At present, even though many assessments may cover the same concepts, the methodological frameworks that underline assessments systems are not the same – meaning that data are not sufficiently consistent to be comparable. By introducing common frameworks for quality in assessments, it will be possible to produce more comparable and reliable data on learning.

The DQAF will cover the institutional environments, statistical processes and data characteristics required for reliable measurement of learning. It will be developed in consultation with stakeholders to ensure its use across different types of assessments and cultural contexts. This will include assessing existing practices of data collection against best practices and internationally-accepted methodologies, which take into consideration of regional contexts and cultural practices. The DQAF is intended to: guide the international education community, regions and countries on data use and priorities for technical assistance; direct country efforts to prepare self-assessment; and allow data users to evaluate data for policy analysis.

3.3. Creation of a platform for input through the Global Alliance to Monitor Learning.

The final pillar is to create a platform for convening and receiving input, and to spur action for measurement in areas of learning that are not yet ready for a universal learning scale. Despite recognition of the importance of global education measurement and consistent support for the UIS mandate from many stakeholders, it is anticipated that some stakeholders may not support the ideas presented here for different reasons, ranging from questions concerning the technical rigor of national and regional assessments to lack of trust in the idea that learning can be summarized using a universal learning scale.

As a UN neutral body, the UIS proposes that the Global Alliance to Monitor Learning (GAML) would serve as the institutional platform to oversee the coordination of efforts to measure learning and the harmonization of related standards.

While cognizant of political and institutional realities, the GAML will also need technical and practical guidance. It will therefore establish a technical standing group (TSG). The criteria underlying the formation and housing of the group should include: impartiality and credibility among peers; technical expertise in cross-cutting issues in the area of learning assessment; and broad geographic and inter-sectoral representation.

Supported by other ad hoc, time-bound technical groups and various specialized agencies and partners, GAML will provide strategic vision and coordination among the different players towards the establishment of a multi-year programme involving methodological and technological innovation, capacity-building and advocacy for learning assessment by engaging different stakeholders in a coherent and coordinated fashion.

This institutional framework, under the aegis of a UN agency and with partners outside of the UN system, is uniquely designed to mainstream learning assessment within larger data efforts and foster integration with other national data sources.

GAML has the following objectives:
1. Establish and promote the implementation of a common international code of practices on learning assessment;
2. Develop and carry out a coordinated programme of methodological work to underpin the development and adoption of improved standards, methods and practices in learning assessment, including the use of data;
3. Strengthen the sustainability of the implementation of learning assessment initiatives in countries.

4. Activities to ensure quality data are mapped within a global metric

The LO monitoring strategy encompasses two main fields of activities, which are presented in Table 1 and discussed in greater detail in Annex A.3.

Table 1: Summary of two main fields of activities of UIS LO monitoring strategy

Activity	Description of activity
1. Development of tools and processes to monitor SDG 4 targets related to learning outcomes. Note: *This is a general field of activity; the precise steps will be defined*	STEPS: 1. Development of a **Global Common Content Framework for Reference (GCCFR)** for each of the learning domains identified in the SDG 4 targets related to learning outcomes. 2. Development of the **Assessment of Data Process (ADP)** for learning outcomes data.

with the start of the work of the LO strategy.	3. Development of a **Global Reporting Metric (GRM)** for each of the learning domains identified in the SDG 4 targets related to learning outcomes.
2. Application of the tools and processes to generate country-level information about progress towards the SDG 4 targets related to learning outcomes. Note: *That this is a recurrent activity; part or all of it will be undertaken on a regular basis, each time country-level learning assessment results need to be used for SDG 4 monitoring.*	STEPS: 1. Map the content of the country-level learning assessment against the content for the learning domain as articulated in the GCCFR. 2. Apply the ADP to determine whether or not the country-level learning assessment is of sufficient quality that its results can be reported against the GRM. 3. Undertake the theoretical/empirical activities required to report the country-level results against the GRM. 4. Report the country-level results against the GRM.

5. Outputs

The strategy will result in the following outputs:

- ✓ A Global Common Content Framework for Reference (GCCFR) for each of the learning domains. The content framework will be used to align and map different national and cross-national assessments contents for each point of measurement

- ✓ A Data Quality Assurance Framework (DQAF) applied to learning outcomes data and its implementation tool the Assessment of Data Process (ADP)

- ✓ A well-defined reporting package

  - o Linkage package (Global Linking Package)

  - o Universal Basic Test (UBT) for each domain and point of measurement

  - o Universal Background questionnaire

  - o Global Reporting Metric including Proficiency levels and related benchmarks

  - o Guidelines for data analysis and data use

- ✓ A developed operational and implementation plan for countries to implement

The Global Linking Package could serve, depending of the coverage of all minimum contents/skills, as the Universal Basic Test (UBT) for each domain and point of measurement. This could be a standalone global public good used both as linking tool between assessment and as a minimum assessment package for countries who do not have yet learning assessments. Countries that administer their own assessment or are joining any Cross-National Assessment (can) could use this as the way to link to the global metric.

Figure below summarizes the current states of measurement in each goal.

Share of countries with a cross-national assessment administered in the last 5 years

	Target 4.1			Target 4.2		Target 4.4	Target 4.6	Target 4.7
	TERCE, PASEC, PILNA, SACMEQ, PIRLS, TIMSS, PISA			EAP-ECDS, MELQO, PRIDI, IDELA, EDI, ECDI, UIS		ICILS, PIAAC	PIAAC, STEP	ICCS
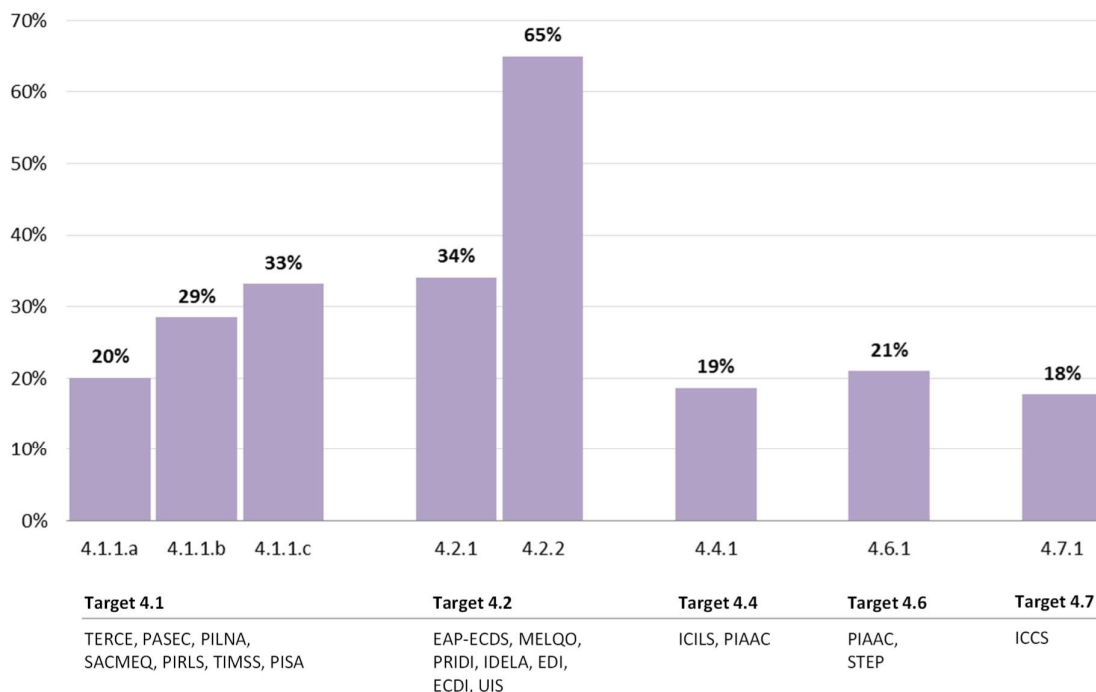
*Given that the national developed assessments in these areas are less known, we are only providing information on the known cross-national assessments.

Source: UIS.

5.1. Prioritization: which education level do we begin with?

Depending on the approach selected by the international education community, the technical process could start with either the primary or the lower secondary education level. From the perspective of test availability, lower secondary is the most logical or easiest level to start with. However, it can also be argued that starting at a lower level would be less costly and more effective to improve learning and reduce school drop-out. Table 2 shows that there are important reasons to start with primary school preferably before the final grade.

Table 2: Out-of-school children and adolescents of primary and lower secondary school age, 2013

	Primary school age		Lower secondary school age	
	Million	Rate	Million	Rate
Arab States	4.9	11.8	3.7	17.0
Central and Eastern Europe	0.7	3.9	0.7	3.8
Central Asia	0.4	6.3	0.5	6.7
East Asia and the Pacific	6.6	4.1	7.7	8.4
Latin America and the Caribbean	4.1	6.5	2.5	6.7
North America and Western Europe	2.2	4.3	0.8	2.5
South and West Asia	**10.3**	**5.9**	**26.2**	**25.6**
Sub-Saharan Africa	**30.1**	**20.5**	**22.7**	**34.5**
World	**59.3**	**9.0**	**64.9**	**17.3**

Source: UIS Data Centre, December 2015

5.2. How can we measure learning levels of out-of-school children and youth?

Given the large out-of-school populations, the global strategy must ensure sufficient measurement of the learning levels of children and youth outside of the formal education system. This could entail generating a short test (based on the universal basic test), which could be incorporated in household surveys (HHS) so that countries can easily measure the level of learning outcomes for out-of-school children. HHS have the advantage of a representative sample and include detailed information on socio-demographic characteristics of the children, which could be particularly useful to identify pockets of children who do not learn within a country. The disadvantage is that not all countries conduct HHS on a regular basis, which would be required for appropriate monitoring. Further, building capacity to run a HHS-based assessment could be even more difficult. A cost-effective solution that could be tested would be to develop an adaptive test based on the Universal Basic Test materials.

Another alternative would be to implement as UBT part of a targeted programme evaluation in a non-representative basis in order to gain a better understanding of specific situations and interventions. The best option will depend on the priorities of each country.

6. Timeline

The timeline is projected on a three-year cycle. Since the development status of each target is different, the targets are grouped as follows: Group1 - 4.1, 4.2, 4.6; Group 2 – 4.4. 4.7.

Table 3: Products by target over the three-year cycle

Target	Products	Year of delivery			
		2016	2017	2018	2019
4.1	1. Technical guidelines 2. International code of practices 3. Learning scale 4. Reporting metric 5. Data warehouse 6. Sustainable capacity plan	Develop DQAF Produce GCCFR and ADP	Develop learning scale and reporting metric Implement linking and conduct mapping exercise Create data warehouse Develop sustainable capacity plan	Implement capacity building plan in countries Policy and technical papers for dissemination	
4.2		Develop DQAF Produce GCCFR and ADP	Develop learning scale and reporting metric Implement linking and conduct mapping exercise Create data warehouse Develop sustainable capacity plan	Implement capacity building plan in countries Policy and technical papers for dissemination	
4.6		Adopt existing DQAF Adopt GCCFR and ADP	Adopt learning scale and reporting metric Implement linking (field test) Update data warehouse Develop sustainable capacity plan	Implement capacity building plan in countries Policy and technical papers for dissemination	
4.4, 4.7			Adopt DQAF	Adopt and/or update	Implement capacity

			Adopt GCCFR and ADP	learning scale and reporting metric	building plan in countries
				Implement linking (field test)	Policy and technical papers for dissemination
				Update data warehouse	
				Develop sustainable capacity plan	

7.

Annex

A.1. Why some assessments are not comparable

Content framework

Depending on the curriculum in a country, national assessments usually have different content coverage for a given grade. Furthermore, even domains can be defined differently. **Tables A-1** and **A-2** present different definitions of literacy and mathematics used in cross-national assessments.

Table A-1: Definition of literacy in various cross-national assessments

PISA 2000	PISA 2009/2015	PIRLS	SACMEQ	STEP
Reading literacy is understanding, using and reflecting on written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society	Reading literacy is understanding, using, reflecting on and **engaging with written texts**, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society	"the ability to understand and use those written language forms required by society and/or valued by the individual".	the ability to understand and use those written language forms required by society and/or valued by the individual	"Understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential"

Source: UIS based on Cresswell, Schwantner and Waters (2015). "Review of component skills assessed and contextual data collection used in relevant international assessments". PISA for Development Expert Paper series. Paris: OECD.

Table A-2: Definition of mathematics in various cross-national assessments

PISA 2015	TIMSS	SACMEQ
Mathematical literacy is an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts	"knowing, applying and reasoning	"The capacity to understand and apply mathematical procedures and make related judgements as an individual and as a member of the wider society"

Source: UIS based on Cresswell, Schwantner and Watters (2015).

Types of items and assessment formats

Assessments can be built in different formats, from multiple choice questions only to a combination of multiple choice and constructed response items *(*see **Table A-3**).

Table A-3: Item development

PISA 2015	LLECE	SACMEQ
▪ Item generation ▪ Panelling items ▪ Cognitive trial ▪ Field trial ▪ Main study selection	Uses the expert group approach in which a group of experts calls for submission of items. TERCE is based on a curriculum analysis, specification tables. Item development involves specialists from almost all countries	Items are developed by a panel of subject specialists drawn from all 15 participating school systems

Source: UIS based on Cresswell, Schwantner and Watters (2015).

Target population

Since assessments can be given at different grades, the target population may vary. Some countries assess at mid-education level, some at the end of an education level, while others at both mid- and end of an education level. Furthermore, the number of years of schooling (or duration of schooling) may vary from country to country.

For example, some countries have six years of primary education so testing at mid- and end of an education level may represent Grades 3 and 6 respectively. Others have four years of primary school so the mid- and end of education level may be Grades 2 and 4 respectively. **Table A-4** shows the average duration of primary and lower secondary education across regions. It is clear how global assessments take place in different grades depending on the end of ISCED 1 and ISCED 2 cycles.

Table A-4: Number of countries/territories by last grade of primary and lower secondary education by region

Region	Last grade of primary education					Last grade of lower secondary education					Total
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	
East Asia & Pacific	1	5	27	2	-	1	7	18	8	1	35
Europe & Central Asia	20	11	18	3	1	-	11	34	7	1	53
Latin America & Caribbean	-	2	30	9	-	-	6	24	11	-	41
Middle East & North Africa	2	4	15	-	-	-	1	18	2	-	21
North America	-	-	3	1	-	-	1	2	1	-	4
South Asia	-	5	1	2	-	-	4	2	1	1	8
Sub-Saharan Africa	-	3	35	10	-	-	6	18	22	2	48
Total	**23**	**30**	**129**	**27**	**1**	**1**	**36**	**116**	**52**	**5**	**210**

Source: UIS.

Data modelling and reporting

In terms of data modelling and reporting, some countries may use more sophisticated models – like item response theory – to scale and report scores, while others may use simple classical theory descriptive statistics, like the proportion of correct scores. Depending on the model used, reporting scores will differ in scale and metrics.

Table A-5: Data modelling and reporting

	PISA	PIRLS/TIMSS	SACMEQ	LLECE
Cognitive reporting scale	Formerly Rasch Model (1-PL for multiple choice items and Partial Credit model for construct response items). Currently data has been rescaled to a 3-PL and Generalized Partial Credit model (GPCM).	2- or 3-PL model for multiple choice items and Generalized Partial Credit model for constructed response item	Rasch model	Rasch model
Performance/proficiency level	6 levels of proficiency with level 6 representing the most advanced level and level 1 representing the basic level.	4 levels of proficiency with level 4 representing the most advanced level and level 1 representing the basic level.	8 levels of proficiency with level 8 representing the most advanced level and level 1 representing the basic level.	5 levels of proficiency with level 4 representing the most advanced level and below 1 representing the basic level.

Source: UIS.

Contextual information

Contextual information is usually collected during national or cross-national assessments through in-school assessments or household assessment surveys. The information can be used to support policy-related analyses of the results and used to improve the design and development of future national and cross-national assessments. The information collected varies across countries and regions, but data collection should include a few common characteristics: gender, grade, age, location, socio-economic background and disability status in order to monitor progress.

Technology

Over the years, technology has enabled a more dynamic assessment design. With improved psychometric modelling, with which reasonable estimations can be done using

a smaller number of items and target populations, different implementation platforms and operational procedures can be used.

For example, the use of computer-based adaptive tests allows respondents to receive a set of test items based on their existing skills so they do not have to sit through a long test. Using computers or tablets as the presentation platform allows a pre-selected set of items to be embedded, scoring of cognitive items and coding of background information to be done automatically. This also improves the efficiency of data processing for further analysis. Furthermore, technology also allows the use of authentic cognitive items, like simulation, to put items into a more realistic situation.

National and cross-national assessments have been progressively updated to incorporate new technology, as seen with the U.S. National Assessment and Educational Progress (NAEP), PISA and PIAAC, to name just a few. Nonetheless, not all assessments implement advanced technology. In addition, there are costs associated with development and assessment design. Therefore, it is important to prioritise the most required features in assessments and budget the development cost accordingly.

Different options to link across assessments

There are various options to link across assessments. Below show a summary of the options and an evaluation index based on technical complexity and cost.

Table A-6: Ranking alternatives: An Evaluation Index

Option	Description	Cost	Technical	Rank
1.Reference benchmark: New asssesment	Develop a new assessment specific to a given target population, limited to a given number of domains, to be implemented in all countries. This option is extreme but offers a wider perspective to evaluate other options.	5	5	5
2.Backward-linking cross-national assessments (CNAs)	Connect results from current assessments using existing frameworks and reporting metrics seeking common denominators. This implies using items and responses from past assessments to build a common denominator.	4	4	4
3.Forward-linking CNAs	Link future cycles of CNAs by redesigning the assessments in a way that makes them comparable. This means agreeing on a core common framework that subsequently will allow to link assessments and reporting metrics.	2	3	2
4.Forward-linking national assessments	This is a natural extension of the forward-linking of CNAs to national levels. Since there are over 200 countries to consider, this a long-term process.	3	4	3
5.Adopt an assessment	Expand or adopt the framework of an existing cross-national assessment taking advantage of coverage in order to include countries or regions currently not covered. This means defining a common core for the assessment that is a sub-set of the current test and including other countries (which may imply more field work)	1	3	1

Source: UIS

A.2. Data Quality Assurance Framework (DQAF) for assessments

The DQAF is used to assess national data quality, covering institutional environments, statistical processes and characteristics of statistical products.

The DQAF provides a structure for assessing the quality of data by evaluating existing practices against best practices and internationally-accepted methodologies. It can be used:

- To guide individuals on data use and set priorities for technical assistance;
- To guide country efforts, e.g. to prepare self-assessments; and
- To guide data users in evaluating data for policy analysis.

A DQAF can help countries identify shortcomings in assessments and update procedures to collect better data for monitoring purposes.

The UIS has developed a DQAF[2] for education data and is working with technical partners to develop one for assessment (please see link in footnote). Most importantly, the DQAF can ease discussions with other international assessment organizations, like IEA, ETS and the OECD, to produce a more elaborate guideline through international collaboration. The following are possible areas of collaboration:

- Harmonise definitions.
- Work on methodological guidelines, for example, on the psychometric principles and applications in the design of an assessment and the use of results.
- Extend DQAF to collect other relevant socio-demographic variables to support statistical development and ensure the use of assessment results in country strategies.

The following are the initial guidelines of the DQAF for assessments:

(1) Prerequisites of quality
   a. Ensure the responsibility for collecting, processing, and disseminating the statistics is clearly specified.
   b. Ensure data sharing and coordination among data-producing agencies.
   c. Ensure respondents' data are kept confidential and used for statistical purposes only.
   d. Ensure statistical reporting through legal mandate.
   e. Monitor the relevance and practical utility of existing data.
   f. Ensure processes are in place to focus on quality.
   g. Ensure procedures are in place to monitor quality during the planning and implementation phase of the program

(2) Assurance of integrity
   a. The statistics produced are not bias.
   b. The process and data are validated through validity studies.

---

[2] The UIS DQAF for education data could be found in the following link:
http://dqaf.uis.unesco.org/index.php?title=The_UIS_Education_DQAF

    c. The data collected and the choice of statistical techniques, as well as decisions about dissemination, are informed by statistical considerations and country strategy.

    d. The terms and conditions under which data are collected, processed, and disseminated are available to the public.

(3) Methodology soundness

    a. The overall structure of assessment in terms of concepts and definitions follows internationally accepted guidelines, or good practices.

    b. The scope of the assessment is consistent with internationally accepted guidelines, or good practices.

(4) Accuracy and reliability

    a. Data are collected taking into account country-specific conditions.

    b. Data collected should follow the defined scope.

    c. Date should be released in a timely manner.

    d. Data collected should be validated via multi-method approach. The data collection and data process should be carefully validated from sampling, administration, scoring, coding, processing, analyses, and estimation. The process of the assessment are carefully monitored and applied overtime.

    e. Potential indicators of problem are investigated.

(5) Consistency

    a. Statistics are consistent within the datasets.

    b. Statistics are consistent and reconcilable with other data sources and statistical procedures.

(6) Accessibility

    a. Statistics are presented in a way that facilitates proper interpretation and meaningful comparisons.

    b. Dissemination media format are adequate.

    c. Statistics are made available to the public.

    d. The levels of detail are adapted to the needs of the intended audience.

A.3. Towards a Global Reporting Metric

Producing quality data for a global metric

The learning outcomes strategy consists of two main activities, presented in Table A-7 and Annex A.2.

Activity 1: Developing tools and processes for monitoring SDG 4 targets related to learning outcomes

Step 1: Developing a Global Common Content Framework for Reference for each learning domain in the SDG 4 targets

There will be one Global Common Content Framework for Reference (GCCFR) for each learning domain that appears in the SDG 4 targets related to learning outcomes.

The GCCFRs will be elaborations of how skills/proficiencies grow within learning domains. The GCCFRs will be developed with a combination of theoretical and empirical activities, as follows:

- **Theoretical activity:** Experts in cognitive development and learning will work together to achieve consensus on substantive descriptions of growth within the learning domains.

- **Empirical activity:** The content of international, regional and perhaps some national learning assessments will be reviewed and mapped in order to get information about content coverage of assessments.

The GCCFRs for each learning domain obtained through these activities will be comprehensive, transparent and coherent.

Step 2: Developing a Assessment of Data Process

There will be one Assessment of Data Process (ADP) that will be applied to all kinds of learning assessments. It will be a process by which the methodologies and products of a learning assessment are inspected and compared against some previously-articulated international standards in learning assessments, with a view to determining whether or not the learning assessment in question is sufficiently aligned with best practices to permit its results to be used for SDG 4 monitoring.

The international standards in learning assessments that are used for comparison will be formulated by the UIS as an activity that is related to but not directly part of its learning outcomes strategy.

Step 3: Developing a Global Reporting Metric for each learning domain in the SDG 4 targets

There will be one Global Reporting Metric (GRM) for each learning domain that appears in the SDG 4 targets related to learning outcomes. The GRMs will be developed by conducting linking exercises through which transformation functions will be obtained. These functions will enable results from one assessment to be

reported on the same scale as results from another assessment, and ultimately enable results from all the involved assessments to be reported on the same scale.

Building a metric could be done using different alternatives.

Table A-7: Alternatives to generate a global reporting metric

		Advantage(s)	Disadvantage(s)
Option 1	Statistical projection	Simple. Database exists. Quick for initial global monitoring.	Might have biased results if assumptions for statistical projection do not hold. On reference to a chosen assessment's reporting metric. Countries might not agree.
Option 2	Link across assessments – linking assessment to reference assessment	Own schedule. Could rely on existing framework of methodologies.	Might have biased results if framework for methodologies varies greatly between linking and reference assessment. If a reference assessment is chosen, all assessments will be referenced to the adopted assessment's reporting metric. Unless a plan is in place to create another reporting metric that is not on reference assessment metric.
Option 3	Link to common item pool	Country could conduct assessment on its own schedule. New reporting metric will be defined so country could do own assessment and yet have its assessment report on a newly defined global reporting metric. Could rely on existing framework for methodologies with adaptations.	Politically challenging because it might be difficult to engage all regional and international assessment bodies to contribute to the item pools although this will be most cost effective. Relatively resource-intensive if new metric is preferred but only this part needs development.

Activity 2: Application of the tools and processes for monitoring of SDG 4 targets related to learning outcomes to country-level learning assessment results

Step 1: Comparing content of a country-level learning assessment to the Global Common Content Framework for References

Before country-level learning assessment results can be considered appropriate for SDG 4 monitoring, it will be necessary to ensure that there is sufficient overlap between the content that is covered in the learning domain(s) in the assessment and the articulation of content for the learning domain(s) that is in the GCCFRs.

If it is determined that there is not sufficient overlap between the content in the learning assessment and the GCCFR, then the assessment will not be included in SDG 4 monitoring at this stage. The country will be advised to adapt its assessment to bring its content more in line with the GCCFRs and supported through this process, if necessary.

If it is determined that there is sufficient overlap between the content in the learning assessment and the GCCFR, then the assessment will proceed to step 2.

Step 2: Applying the Assessment of Data Process to a country-level learning assessment

The country-level learning assessment methodologies and products will be judged against the Assessment of Data Process (ADP) to determine whether the assessment results are acceptable for SDG 4 monitoring.

If the country-level assessment does not pass the ADP, then the results will not be used for SDG 4 monitoring. The country will be advised to adapt its assessment to bring its methodologies and products more in line with the standards in the ADP and supported through this process, if necessary.

Step 3: Undertaking theoretical/empirical activities required to report country-level assessment results against the Global Reporting Metrics

Once a country-level assessment passes through the two-stage validation, the data will be processed and put into a central data bank to be used for calibration. The data will be aligned to the common scale and transformed into the global reporting metric.

Step 4: Reporting country-level learning assessment results against Global Reporting Metrics

Once the country-level assessment data have been transformed onto the global reporting metric, it will be put onto a platform. Data will be available to countries and researchers for further analyses. **Figure A-1** summarises the validation process.

Figure A-1: Validation process towards Global Reporting Metrics
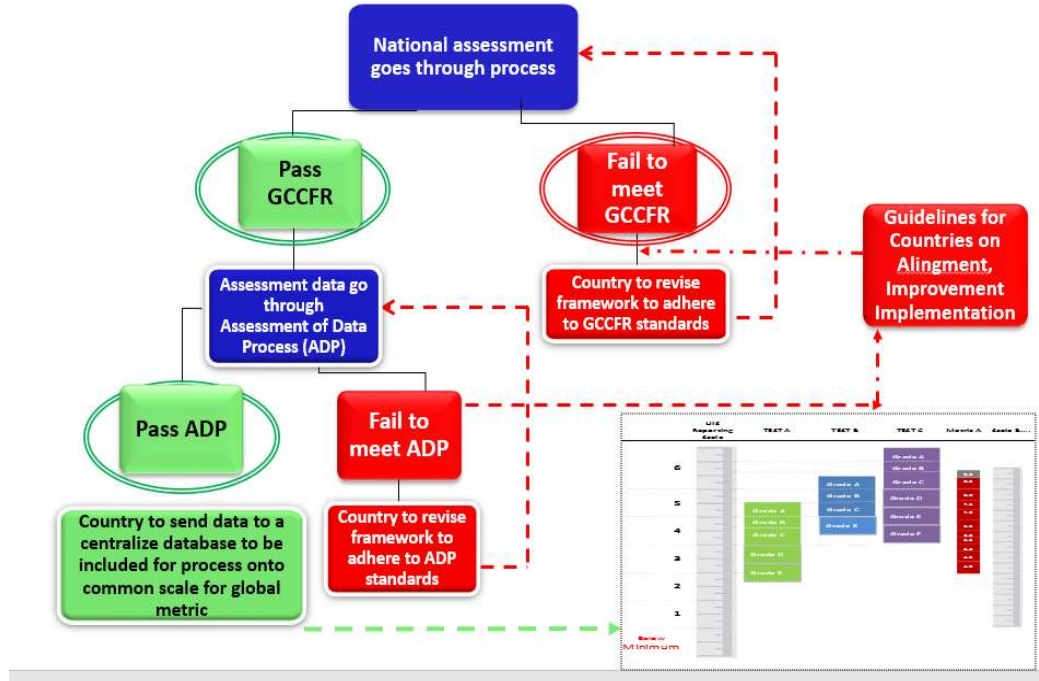
A.4. Hanushek-Wößmann's methodology of a common metric

Hanushek and Wößmann's estimation of a common score from a variety of cognitive achievement tests does not take into account the different objectives and targeted samples which participated in the assessments. In addition, there are some assumptions that might not be applicable when countries at different levels of development are involved. The common score is derived to be used as a vehicle to estimate growth and macroeconomic performance across countries which have participated in international large-scale assessments, the interest of having a 'rank order' score does not provide us with comprehensive information to help policymakers to develop plans for its population's learning.

Assessments used

The analysis relies on a variety of cognitive achievement tests: information from the international assessments include a set of countries voluntarily participating in the International Association for the Evaluation of Educational Achievement's (IEA) TIMSS (and earlier versions of mathematics and science assessments) and in the Organisation for Economic Co-operation and Development's (OECD) PISA.

These tests have different groups of countries, samplings of students, and perspectives on what should be tested (see Neidorf, Binkley, Gattis, and Nohara (2006)). Hanushek and Wößmann's approach is to aggregate across the variety of tests for each country in order to develop a composite measure of performance.

However, the tests do not measure a common dimension of cognitive skills. The TIMSS mathematics and science tests are developed by an international panel but are related to common elements of primary and secondary school curricula, while the PISA tests are designed to be assessments of more applied ideas. In their development of a common metric, they also employed data from the U.S. National Assessment of Educational Progress (NAEP). NAEP, which is conceptually closest to the TIMSS tests – except that it relates more directly to U.S. curriculum – provides information over time on a consistent basis.

Part of the analysis on individual returns relied on the International Adult Literacy Survey (IALS), a set of tests given to 20 countries between 1994 and 1998. These tests cover several functional areas: Prose Literacy – the knowledge and skills needed to understand and use information; Document Literacy – the knowledge and skills required to locate and use information contained in various formats; and Quantitative Literacy – the knowledge and skills required to apply arithmetic operations. They were designed to be very practical.

Observations

The TIMSS tests with their curricular focus and the PISA tests with their real-world applications are highly correlated at the country level. For example:

a) the correlation coefficients between the TIMSS 2003 tests of 8th graders and PISA 2003 tests of 15-year-olds across 19 countries participating in both are 0.87 in mathematics and 0.97 in science.

b) they are 0.86 in both mathematics and science across 21 countries participating both in the TIMSS 1999 tests and the PISA 2000/02 tests.

c) similarly, there is a high correlation at the country level between TIMSS' curriculum-based tests and IALS' practical literacy adult examinations (Hanushek and Zhang, 2006).

According to them, tests with very different foci and perspectives tend to be highly related, lending support to their approach of aggregating different tests for each country.

The general idea behind their approach to aggregation is that of empirical calibration. They rely on information about the overall distribution of scores on each test to compare national responses.

This contrasts with the psychometric approach to scaling that calls for calibrating tests through the use of common elements on each test, i.e. common items to link tests. In reality, each of the testing situations is a separate activity with no attempt to provide common scaling.

They claimed that the strength of their approach is that different tests across a common subject matter are highly correlated at both the individual and aggregate levels. Thus, the distributional information that they used is closely related to variations in individual performance levels. However, this remains an assumption to be tested.

Methodology to create a common metric

Hanushek and Wößmann used data from international student achievement tests on 12 international testing occasions. Including separate tests in different subjects and at different age groups, these testing occasions yield 36 separate test observations altogether, each with between 11 and 45 participating countries with internationally-comparable performance data.

Most of the tests were conducted by the IEA, with the exception of the OECD-conducted PISA tests[3]. In order to make performance on the different international

---

[3] In this study, Hanushek and Wößmann do not include the two tests conducted by the International Assessment of Educational Progress (IAEP) in 1988 and 1991, because they used the U.S. NAEP test as their testing instrument, which is geared to the U.S. curriculum and may thus introduce bias to the international testing. By contrast, the tests included here are not associated with the curriculum in any particular country but have been devised in an international cooperative process between all participating countries.

tests comparable, Hanushek and Wößmann (2015) developed a common metric to adjust both the level of test performance and the variation of test performance through two data transformations.

d) First, because the United States has both participated in all of the international tests and has maintained its own longitudinal testing (the NAEP), Hanushek and Wößmann (2015) calibrated the U.S. international performance over time to the external standard – thus benchmarking each of the separate international tests to a comparable level.

e) Second, while this provides a relative comparison of countries taking each test over time, it is also necessary to establish the variance of the tests so that direct compatibility of countries taking different tests can be established. The calibration of the dispersion of the tests relies on holding the score variance constant within a group of countries with stable education systems (defined in terms of secondary school attendance rates) over time. For this, Hanushek and Wößmann (2015) used the 13 OECD countries who had one-half or more students completing upper secondary education around the beginning of international testing in the 1970s as the "stable" country group, and standardised variances to their group performance on the 2000 PISA tests. The details of the transformation are found in Hanushek and Wößmann (2015).

Limitations

Given that they used data on education systems in OECD countries as the basis to calibrate the score variance of other countries and defined stable education systems in terms of secondary school attendance rates, these assumptions might not be applicable to developing countries with varied education systems. A more direct calibration psychometric approach should be considered which means that more investment in data collection and funding proposition are required.

Furthermore, the purpose of their development of common metric across international assessment is to look at economic growth, a rank order based on correlation across assessments used may be sufficient for the purpose of their analysis. As for the UIS, the interest is on comprehensive information of education system, learning of targeted population including contextual data for policy development. Due to the difference in purpose, the way to generate common metric for analyses should be difference. The way they generate common metric might be sufficient for their analyses purpose but it might not be sufficient for the SDG global monitoring purpose.

A.5. Indicative budget

Below two tables show the indicative cost of the five SDG targets over the three-year period and the estimated cost by function over the three-year period.

Table A-8. Estimated cost for each of the 5 SDG 4 targets 2016-2018 (in USD)

Total cost by SDG Target over 3 years

Target	Cost in USD
4.1	3,064,000
4.2	1,155,000
4.4	1,155,000
4.6	1,417,000
4.7	1,155,000
GAML coordination	2,430,000
TOTAL	**10,376,000**

Table A.9. GAML summary budget, 2016-2018

Costs in USD	2016	2017	2018	Total
Management costs (coordination and dissemination)	810,000	810,000	810,000	2,430,000
Develop reporting scales and metrics (standards and guidelines)	1,191,900	1,191,900	1,191,900	3,575,700
Build data warehouse and system mapping (of learning assessments)	1,456,767	1,456,767	1,456,767	4,370,300
Total	**3,458,667**	**3,458,667**	**3,458,667**	**10,376,000**

Reference

Hanushek, Eric A., and Wößmann, Ludger (2015). The knowledge capital of nations: education and the economics of growth. MIT press ebooks.

Hanushek, Eric A., and Lei Zhang. 2006. "Quality Consistent Estimates of International Returns toSkill." National Bureau of Economic Research, WP12664, Cambridge, MA, NBER November.

Neidorf, Teresa S., Marilyn Binkley, Kim Gattis, and David Nohara. 2006. *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments*. Washington: National Center for Education Statistics (May).