United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

TECHNICAL COOPERATION GROUP

SUSTAINABLE DEVELOPMENT GOALS

TCG4:
SDG 4 Benchmarking: Background paper

TCG4/33

16-18 January 2018
Dusit Thani Dubai
133, Sheikh Zayed Road, Trade Centre,
Dubai, United Arab Emirates

Contents

Abstract

Monitoring progress towards SDG 4 is critical on many levels for international educational development. Each of the 7 targets and 3 means of implementation for SDG 4 contain specific objectives that countries should collectively achieve by 2030. However, there is no concrete clarity yet on whether the performance of individual countries/regions will be measured against pre-defined standards over time until 2030. In order to shed some helpful insight on this topic and to explore the possibility of establishing benchmarking for SDG 4, this study takes an exploratory approach and reviews existing ideas and research on benchmarking of the SDGs and similar goals. The paper starts by providing a brief background to the concept of benchmarking and views on the need for the development of SGD 4 benchmarks to set the context. The results from literature and documents review are organized around three major areas: (a) existing recommendations and views on SDGs benchmark development expressed by agencies and scholars in the field, with a focus on SDG 4; (b) existing benchmark initiatives in education in the context of the European Union, OECD and developing regions; and (c) strategies and benchmark options adopted or in discussion for other SDGs. Finally, a preliminary conceptual framework and initial pointers and recommendations are presented in relation to SDG 4 benchmarking efforts.

Introduction

The depth and breadth of the 17 Sustainable Development Goals (SDGs), adopted in September 2015, represent the bold vision and aspirations of the global community for 2030. Education, under SDG 4, focuses on ensuring inclusivity, equitability, and quality within a lifelong learning perspective. With its seven targets and three means of implementation, the fourth SDG, which cuts across several other goals, shows a level of ambition for the next 15 years that goes beyond any previous global education agreement (UNESCO, 2016a). (See Appendix for a detailed list of the targets.)

Monitoring progress towards SDG 4 is critical to show what needs to be done, by when and by whom. This has been outlined in the Incheon Declaration and the Education 2030 Framework for Action to monitor and report on SDG 4 and on education in the other SDGs (UNESCO, 2016b). The ambition of SDG 4 is reflected in its focus on improvements in education outcomes, such as learning achievement; access at post-basic education levels, including upper secondary and tertiary education; its lifelong learning perspective, including adult education; the reduction of disparity in education based on factors such as wealth, gender and location; and changes in the content of education to better align it with the challenges of sustainable development. There is a huge task ahead for national statistical systems to monitor progress towards SDG 4 and for education ministries to make effective use of the new information (UNESCO, 2016a). As national education systems vary in terms of structure and curricular content, it can be difficult to benchmark performance across countries over time or monitor progress towards national and international goals. In order to understand and properly interpret the inputs, processes and outcomes of education systems from a global perspective, it is vital to ensure that data are comparable (UIS, 2012).

While global monitoring is not the main objective of the Education 2030 Agenda, its contribution should not be underestimated. Its purpose is not to impose particular global norms, but rather to facilitate well-informed and meaningful comparisons between countries and regions, spurring debates, especially between governments, civil society and engaged citizens, for the improvement of the educational systems (UNESCO, 2016a). Each of the 7 targets and 3 means of implementation for

SDG 4 contain specific objectives that countries should collectively achieve by 2030. However, there is no concrete clarity yet on whether the performance of individual countries/regions will be measured against pre-defined standards over time until 2030. Therefore, to better inform this topic and to explore the possibility of establishing benchmarking for SDG 4, this study takes an exploratory approach and reviews existing ideas and research on benchmarking of the SDGs and similar goals.

The rest of the paper is organized as follows. The first section provides a brief background to the concept of benchmarking and views on the need for the development of SGD 4 benchmarks to set the context. The second section discusses the method used in the literature review. The third section presents results from the literature review that is organized around three major areas as discussed in the methods section below. The fourth section provides a preliminary proposal of a conceptual framework for benchmarking. The fifth section summarizes and concludes.

Historical Background: Benchmarking in Global Educational Goals

Originally developed in private sector management, benchmarking now features prominently in the public sector to improve the efficiency, effectiveness, and legitimacy of government and institutional interventions (Broomé & Quirk, 2015; Groenendijk, 2009). Benchmarking in education occurs when measurable standards are set for learning, the definition of which has evolved to focus on learning outcomes rather than just access and parity (World Bank, 2018). Educators and governments are additionally paying increasing attention to international comparisons as they seek to develop effective policies to improve the performance of their education systems. Nations and states are therefore working to benchmark their education systems to establish a solid foundation for economic development in the 21st century.

Clemens (2004) argues that the development goals, since their inception after World War II, have been based on the simple idea that schooling children produces an educated workforce and thereby increases national productivity and income. With evidence generally supporting this, as well as additional factors such as family income and parental education levels significantly influencing learners' education outcomes, development goals have reinforced and are reinforced by the belief that educational attainment is mostly supply-constrained (Clemens, 2004). Fueling the human capital resource has therefore become the most important focus in the new global economy (King, 2015). As a result, measurement for development goals has typically and naturally relied on economic input (such as GDP) as the influencing factor in achieving them. This is exemplified in the World Bank's emphasis of rates of return analysis, which has become the most important analytical tool to guide its education policy since the eighties (Verger et al., 2012).

With the post-1970s era came the weakening of the notion that education systems could not be compared with and to one another (Kamens, 2013). Since then, international comparison, with the pursuit of discovering the best practices that produce high educational outcomes, has become the *de facto* influence for both legitimacy and funding. Along with the push for comparative assessment, development goals increasingly contribute to this pressure of international benchmarking for assessing both national and world educational progress. As Kamens (2013) writes, this has created an emerging "horse race mentality" about educational progress and success, which has in turn created and legitimized an international "audit culture" (p. 117).

With the development goals initiative well into more than sixty years of implementation and iterations, the ways in which countries have been benchmarked typically fall into two types. According to Vandemoortele & Delamonica (2010), benchmarks can measure performance in either absolute or relative terms. During the MDG era, most of its goals were expressed in relative terms; for example, reducing poverty by half and infant mortality by two-thirds. They add that the inverse relationship between proportional change and the initial situations of countries tended to put the MDG targets in a one-size-fits-all box, which naturally disadvantaged low-income countries (p. 15). Additionally, global goals and targets were earlier expressed in either absolute terms or as combined relative and absolute benchmarks. The SDGs have mostly adopted absolute benchmarks (e.g., achieving an under-five mortality rate of 25 per 1,000 live births). Neither type of benchmark taken alone, however, provides the full picture of a country's progress or situation (van Bergeijk & van der Hoeven, 2017).

The evolution of the goals and how they were monitored underwent further changes. The 1990 World Conference on Education for All (EFA) in Jomtien produced a framework for action established with education targets, but at only the national level. Countries were encouraged to set clear objectives and measureable targets for the decade based on six proposed themes by UNESCO: early childhood, universal primary, learning achievement, adult literacy, essential skills, and skills and knowledge via media (UNESCO, 1990). However, the push for a timeline during the EFA conference began the tension between global goals and targets and national ones that later resurfaced in further global goal-setting initiatives such as that during the Incheon World Education Forum (King, 2015). Until the MDGs and six EFA Dakar Goals in the early 2000s, any efforts of large-scale targets lacked a regular monitoring system.

As international stakeholders discussed the post-2015 agenda, it "became progressively clearer" that education be fully integrated into the 2030 SDGs while at the same time continuing the wider coverage of the EFA Goals (King, 2015: 19). This emphasis resulted in SDG 4 encompassing a wider range of the factors that constitute learning, yet facing greater challenges in measuring what some term as unmeasurable. Because the complexity of the SDGs far surpasses that of the MDGs, the need for a more holistic approach in terms of international standards against which to hold countries has been emphasized. Subsequent discussions and declarations, including the Muscat Agreement in 2014 and the Incheon World Education Forum, emphasized the importance of setting targets at the national/local level but maintain a universal narrative, as well as the responsibility and applicability of the goals for developed countries, to accommodate this approach.

Definitions of Benchmarking

Definitions matter. For example, it is commonly cited that approximately 250 million children do not have basic literacy and numeracy skills in the world. Though this number in and of itself is an extrapolation of multiple estimates, the exact number depends upon which definition is used (UNICEF, 2017). In order to develop a global benchmarking system, definitions take on the upmost importance. Global benchmarking has been generally defined as a range of techniques, such as audits, rankings, indices, or baselines, that can systematically assess performance (Broomé & Quirk, 2015). Table 1 provides a helpful summary of different terminology for use as a reference point of some terms employed in discussions of comparison (provided by Chakroun & Ananiadou, 2017).

Groenendijk (2009) indicates that benchmarking can be done in various ways. First, benchmarking can be internal or external. Internal benchmarking involves units or sub-units within the same entity,

either with the objective to improve (benchmark as learning) or as a means of hierarchical control (disciplinary benchmarking). The distinction between bottom-up and top-down benchmarking is also linked to these two objectives. External benchmarking aims compare entities with the same or very similar characteristics and can be either competitive or non-competitive, with the latter focusing on learning from others.

Benchmarking can also be functional or generic (Groenendijk, 2009). Assuming that entities are comparable, functional benchmarking sets out to analyze aspects such as functions and processes of participating entities independently of characteristics like output and sector, whereas generic benchmarking involves all aspects of involved entities. Standards benchmarking refers to setting a standard of performance that an entity is or could be expected to achieve. With the various ways of benchmarking, careful consideration of the type of benchmarking used must be made to ensure that organizations and/or countries are seen as cooperative partners rather than competitive adversaries (Groenendijk, 2009).

Standard setting is another close or synonymous term with benchmarking. It is defined in the international assessment sphere as "the procedure of defining frameworks for different performance levels, identifying cut-scores on the score scale defining the threshold between levels, and developing substantive descriptions of what the students classified into any specific level are able to do" (Treviño & Ordenes, 2017: 5). Cut-scores are known as the knowledge and skills evaluated in a specific assessment, while levels are the definition of the different categories of performance developed in that assessment (see Table 2 for a list of achievement levels from cross-national assessments). How cut-scores are determined in the assessment procedures will directly establish the definition of the levels of performance. Like qualifications frameworks, these achievement levels will mark the skills and knowledge that learners should have mastered by reaching a particular level, and have direct implications on establishing a global definition of minimum proficiency.

Table 1: Summary of terminology related to reference process

TERM	DEFINITION	ACTION/OUTCOME CONTINUUM
Alignment	Agreement, alliance or cooperation among persons, groups.	Implies a political agreement (e.g. Shorter Oxford third edition – fall into line with)
Benchmarking (study)	Test or measure something against a standard (or develop the standard) UAE: means the continuous process of measuring and comparing products, services and practices with comparable systems or organisations both inside and outside the UAE for the purpose of continuous improvement	Implies an unequal relationship: one side has set the standard against which the other is measured
Comparability (study)	Capable of being compared; features in common	HK/EQF: analyses the technical and conceptual characteristics of the respective frameworks in the context for which they are designed and, by comparing the two frameworks, seeks to identify key aspects of similarities and differences and thereby determine the comparability between them.
Comparative (analysis)	Using comparison as a method of study	Designating the degree of comparison (Shorter Oxford third edition)
Compare	Liken, similar to	Consider or estimate the similarity or dissimilarity of one thing to another (Shorter Oxford third edition)
Comparison	Considering the common characteristics between two or more 'things'	Action or an act of likening or representing as similar (Shorter Oxford third edition)
Compatible	Used to establish a system-to-system level agree ability or harmony of national qualifications frameworks level descriptors and qualifications definitions (New Zealand glossary)	
Compatibility	Consistent with something else, agreement, correspond, in accord	Used in NZ/Ireland report 2010
Mapping	Delineation, representation, reflection	Lower level type of comparison
Reference point	Basis or standard for evaluation, assessment or comparison (Shorter Oxford third edition)	Used in EQF terminology
Referencing (process)	Process that results in the establishment of a relationship between the levels of local, national or regional qualifications framework) (EC, 2013 p. 6).	EU context is a political requirement
Relationship (analysis)	Connection, association, involvement between parties; state or mode of being related/connected (Shorter Oxford third edition)	
Relative (to)	Reference to, relating to (Shorter Oxford third edition)	Thesaurus: comparable, related

Source: Booker 2016 (as cited in Chakroun & Ananiado, 2017)

Table 2. Achievement levels in cross-national assessments

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

Source: Treviño & Ordenes, 2017.

The literature suggests that there is no fixed set of steps to be conducted in standards setting. However, Hambleton (1998) provides a generic set of steps in standards-setting exercises that are common to the process of setting performance standards. Similar adoptions are seen in works of Cizek (2006) and Hambleton et al. (2012). These generic steps can be summarized as below (UIS, 2017e):

1. Choosing a standard-setting procedure including the necessary preparations.

2. Selecting a large and representative panel of experts.

3. Preparing descriptions of the referent population as well as of the performance categories.

4. Training participants to use the standard-setting method.

5. Compiling item judgements/ratings from experts and summarizing outcomes to provide feedback on ratings.

6. Bringing together the experts for discussion based on the feedback from the rating exercise.

7. Experts revising their ratings on the basis of the feedback and discussion

8. Experts finalizing their ratings to determine a recommended standard.

9. Evaluating the process with the participants to get their confirmations

10. Gathering relevant documents pertaining to the validity of the standards being formed.

Currently, global benchmarking definitions remain elusive and complicated. For example, having a clear definition of what constitutes minimum proficiency would require careful thought as to what the international community believes are the concrete knowledge and skills that a learner needs in order to participate in society in a competent way. Treviño & Ordenes (2017) state that such skills must incorporate, among others, the ability to "exert citizenship" and "have the tools to conduct the personal project of life" (p. 24). What this means at a national level, notwithstanding at the global level, has both differing definitions and unclear implications. Further, the word "quality" also features prominently in SDG 4's goals and targets, but what it might mean in the various levels of learning has yet to be defined. Depending on the organization, field, or even theory, even the definition of benchmarking itself has been at the center of many debates.

Who defines benchmarks also matters. International Political Economy (IPE) theories put their emphasis on economic factors as the main drivers of educational change. For IPE theories, international governmental organizations (IOs) are viewed as key transmitters of particular views of education and educational reform to national contexts. Roger Dale (1999, as cited in Verger et al., 2012) systematizes a range of policy mechanisms activated by IOs and other external actors that allows them to frame and influence national and sub-national education policies. One such mechanism is standardization, which is defined as the international community defining and promoting the adhesion to a set of policy principles and standards that frame the countries' behavior (e.g., international performance tests, such as PISA, contribute to the standardization of curricular content at the global level) (NGA, 2008).

However, benchmarking at the international level runs into problems in not only cooperation in defining those principles, but also choosing the criteria and indicators with which to monitor them (Groenendijk, 2009). Because policies cannot be easily defined in terms of input and output, and both national and international trends affect performance, finding indicators that are easy to interpret yet meaningful, consistent, and sensitive to complexity is the holy grail in comparing countries internationally. The SDGs, as will be discussed in a later section, have taken into account country context in its targets framework, but it is this variety of needs and education realities – which are directly related to the level of development of each nation – that problematize the effort to define a minimal level of competency for SDG 4.

Recent Developments Seeking Benchmarking Options

Following the adoption of the 2030 Agenda for Sustainable Development and the Education 2030 Framework for Action in 2015, the UNESCO Institute for Statistics (UIS) has been leading the development of a thematic indicator framework for the follow-up and review of SDG 4 on education. A set of 43 indicators, including the 11 global indicators recommended by the Inter-Agency and Expert Group on SDG indicators (IAEG-SDGs), were approved in October 2016 by the Technical Cooperation Group for SDG4-Education 2030 Indicators (TCG). There is one indicator per target

except in the case of Target 4.2, for which two are proposed. Four are identified as Tier I indicators ('established methodology … and data regularly produced by countries'), three as Tier II indicators ('established methodology … but data are not regularly produced by countries'), two as Tier III indicators ('no established methodology') and two have been classified at multiple levels (IAEG-SDGs, 2016).

The TCG mimics some of the features of the IAEG-SDGs. For example, it includes the 28 countries of the IAEG-SDGs as members. In addition, at its first meeting in May 2016, the TCG discussed a classification of indicators like the one used by the IAEG-SDGs. As part of this process, it identified eight Tier III indicators that need further work, either because they are not sufficiently aligned with the concept or because implementation challenges are envisaged (UIS, 2016b).

Method

In this review, we address three questions/essential areas of focus:

A review of existing recommendations and views on SDGs (with focus on SDG 4) benchmark development expressed by agencies and scholars in the field,

A review of existing benchmark initiatives in education in the context of the European Union, OECD and developing regions,

(c) Strategies and benchmark options adopted or in discussion for other SDGs.

To conduct this literature review, we adopted the following method. First, as per feedback from the experts in the field and our own understanding of the subject, we finalized three areas for the review as noted above. Second, we collected papers, briefs, books, and reports that relate to benchmarking, particularly relating to the international assessment and SDGs. Once we had a list of these resources, we examined all the citations used in those studies and then compiled a list of studies that were relevant to the use of benchmarking research. In addition, we cross-examined the articles for relevant citations. Third, we conducted a comprehensive search for papers, articles, and reports on benchmarking on international educational goals on Google Scholar and other search engines (such as Penn Library's Franklin/Article+) that were published till date. This search was conducted for (a) terms such as 'benchmarking' and 'standardization' and 'framework' (b) benchmarking of SDG Goal 4, (c) benchmarking of SDGs in general, and (d) other related areas of benchmarking/standardization in global education. We then analyzed these resources as they related to the designated areas of, pulling out main themes and relevant data that we then highlighted in the results section below.

Results

I. Existing views and recommendations on SDGs benchmark development

This section presents views and recommendations as well as concerns and challenges relating to SDGs benchmarking, with focus on SDG 4. A summary of these views/alternatives along with their advantages and disadvantages is included in Table 5 at the end of this section.

A. Views on SDG 4 Benchmarking

Synthesis Report of the UN Secretary-General of the Post-2015 Agenda

The Synthesis Report of the UN Secretary-General on the Post-2015 Agenda, released in December 2014, offers a useful framework to understand the layers of monitoring required (United Nations, 2014). The report identified four levels, each of which has distinct implications for indicator selection:

Global: To monitor the 17 goals and 169 targets, globally comparable indicators are needed. Countries would commit to report on them and the results would appear in an annual SDG Report that would succeed the MDG Report.

Thematic: The scope of a set of global indicators that aims to capture the entire development agenda will be unlikely to fully satisfy the needs of communities interested in specific goals and themes. An additional set of globally comparable indicators is therefore needed for individual targets within goals such as education.

Regional: Some indicators may not be globally relevant but are essential for regional constituencies to respond to specific contexts and policy priorities.

National: Every country has its own context and priorities, which call for tailored monitoring and reporting mechanisms.

The Open Working Group on the SDGs (OWG) proposal argues that the goals "constitute an integrated, indivisible set of global priorities for sustainable development" (United Nations, 2014). It underscores, however, the importance of each government setting its own national targets inspired by the global level due to country context. This condition is accounted for in SDG 4 with the inclusion of national percentages in four of the education targets and means of implementations (though, without many in other OWG targets) (King, 2015).

A set of Global Reporting Indicators for the SDGs is required to ensure coordinated monitoring and knowledge-sharing at an international level, as well as to support national efforts to measure the SDGs (de la Mothe et al., 2015). These indicators will derive their definitions from official data sources such as censuses, but also from specialist agencies, especially for those areas where data neither exists nor is yet operationalized. The Sustainable Development Solutions Network (SDSN) recommends that each Global Reporting Indicator has at least one lead technical or specialist agency, as well as have a maximum number of 100 global indicators to accommodate country capacity in monitoring them (de la Mothe et al., 2015).

Some concerns about the global framework and the indicators it has or lacks speak to the gaps in the 2030 Agenda in regard to common but differentiated responsibilities of countries and reducing inequalities. While the global indicators are intended for global follow up and review and not necessarily applicable to all national and regional contexts, some argue that this position nevertheless reflects a pressure on all countries to use the global framework as the starting point, risking the exacerbation of its weaknesses without maintaining its strengths (Adam & Judd, 2016). Despite the more participatory nature of the SDGs, the discourse has been heavily driven by the Global North, and concerns that the principal focus of the SDGs is on developing countries. King (2015) notes that every SDG, save for SDG 5, has one or more targets or means of implementation expressed in terms more specifically for developing countries, while developed countries are not as much the focus. It begs the question of whose interests and priorities, in addition to global concerns, took precedent over how easily a target could be monitored.

Turning to thematic reporting, coordination among specialized organizations, universities, and even businesses, who all may have access to data that monitor the thematic indicators as highlighted by the OWG, is also required. For example, the International Fertilizer Association can assist in monitoring SDG 15 with its comprehensive database (de la Mothe et al., 2015). Like global reporting, it may be necessary to have a lead agency for different themes, convening multi-stakeholder meetings to compile reports and innovation in data and metrics.

For regional reporting, existing mechanisms, such as the Regional Economic Commissions, should work as a foundation to foster dialogue and knowledge-sharing among similar regions. Regional monitoring processes can also negotiate what is being measured at the national and global levels, especially if organizations are already subsidiaries of international organizations.

At the national level, the brunt of the work will fall onto National Statistical Offices, though other stakeholders must be involved to mirror the breadth of the SDG agenda (de la Mothe et al., 2015). Ownership of the SDGs at this level has been emphasized and is a crucial component to this process, which means reporting will respond to the priorities and needs of each participating country. As such, nations may choose a combination of the Global Reporting Indicators and the Complementary National Indicators to harmonize global and national reporting. Ensuring that countries have accessible, comprehensive and communicable data can enhance the monitoring of progress within the SDGs at the local and subnational levels of government.

Incheon Declaration

Following the recommendations by the Synthesis Report, the Incheon World Education forum in May of 2015 produced a Declaration stipulating that governments are expected to translate global targets into achievable national targets based on their education priorities, national development strategies and plans, the ways their education systems are organized, their institutional capacity, and the availability of resources. This requires establishing appropriate intermediate benchmarks (e.g., for 2020 and 2025) through an inclusive process, with full transparency and accountability, engaging all partners so there is country ownership and common understanding. Intermediate benchmarks can be set for each target to serve as quantitative goalposts for review of global progress vis-à-vis the longer term goals. Such benchmarks should build on existing reporting mechanisms, as appropriate. Intermediate benchmarks are indispensable for addressing the accountability deficit associated with longer-term targets (UNESCO, 2016b).

Missing from the Declaration, as well as the final version of the UN General Assembly's 2030 Agenda, however, are percentages in key education targets. Instead of a sweeping "all adults" achieving literacy skills as previously discussed, there is now a vague "substantial proportion of adults" (United Nations, 2015, p. 14). The term substantial has also replaced all in the targets for qualified teachers and the number of youth and adults who have relevant skills. This may be viewed as a screen behind which countries can hide less than ideal results in their SDG reporting, magnifying concerns that the indicators and targets miss key data from marginalized communities (Adams & Judd, 2016). The implications for such ambiguity must be explored further.

OECD

The Organization of Economic Cooperation and Development (OECD) has constructed their own benchmarking to monitor their member states' progress towards SDG 4 targets and claims to have

the most comprehensive international benchmarks (OECD, 2017a). For example, they use PISA level 2 in both mathematics and readings to measure the minimum proficiency level in science, mathematics and reading. As per OECD, "there also seems to be a considerable progress to be made on what are classified as 'means of implementation' targets (Targets 4.a, 4.b and 4.c) – those which are meant to guarantee the essential structure and resources needed to achieve all other SDG 4 targets. Among these, OECD and partner countries must work to continuously improve student well-being and the quality of the teaching profession" (OECD, 2017a).

Minimum proficiency levels are established in relation to the OECD average. In reading, this threshold is defined as being able to read to using reading for learning, while in mathematics, it involves a basic understanding of fundamental mathematical concepts and operations (OECD, 2016b). It is important to note that though the OECD average is commonly used as a reference for comparison, the situation in any given country or economy may differ greatly from the average, and sometimes the very low performers on PISA may include students who perform well relative to other student in their country/economy (OECD, 2016b). It is also possible that a low-performing student in PISA may also be considered a high-performing student on a different assessment.

In a recent study on OECD countries' current performances on the SDGs, the organization identified 131 indicators covering 98 targets spanning all 17 Goals that could be measured by OECD data (OECD, 2017c). It notes, however, that its data cannot cover 57% of the targets to be evaluated, as well as the reality that significant statistical work is needed to fill some of these gaps.

Building on the UN global indicator framework, OECD's assessment relied on a dataset that measures countries' relative distances from SDG targets, with available data. Among its indicators are 65 directly comparable ones from OECD's database (e.g., productivity growth), while 14 OECD-sourced indicators served as proxies for those indicators without available data. Further, 37 indicators from the UN Global Indicators Database, like the prevalence of moderate or severe food insecurity, are used where no OECD data exist. Finally, 15 OECD indicators were used that were not on the IEAG Global List, such as on that measures social assistance adequacy. The indicators selected have face validity, discriminatory power, broad availability, and high statistical quality. OECD indicators in this study were able to cover about 100% of SDG 4 (OECD, 2017c). Some of the indicators include the participation rate of youth and adults in formal and non-formal education from PIACC (Goal 4.3.1) and the proportion of schools with access to computers for pedagogical purposes and the percentage of 15-year-olds with access to computers for educational purposes (Goal 4.a.1) via PISA data (OECD, 2017c).

From there, the study examined the distance to travel in order to reach each target level, which involved determining levels of achievement on each target level. At times, the level was pre-determined in the 2030 Agenda, either as a fixed value or as a relative improvement on a country's starting position. International agreements on other relevant targets were used to specify the fixed or relative value. For those values where no international or explicit agreement is determined, currently covering 36 indicators, the study set the best value at the "90th percentile," meaning the level which only 10% of OECD countries currently achieve (OECD 2017c; See Table 3 for further detail).

The study found significant variation among countries' distance to achieve the SDG goals and individual targets, as well as the variation in data coverage. From this, OECD suggests that national

priorities for implementing the SDG agenda should be set at target level, rather than at the goal level. OECD also suggests that, to implement the 2030 Agenda, countries may need to develop additional indicators and evidence to identify and track progress on policies that drive outcomes at the country level and that have significant trans-boundary impacts. Other recommendations suggest that the future statistical agenda on SDGs will have to increasingly concentrate on policy levers and global contributions. For the latter, it will be important to identify spill-overs from domestic policies contemplated in the Agenda 2030.

Table 3: Types of SDG indicators and their 2030 end-values

Type of indicator	Means of setting 2030 end-value	Number of indicators
A1. SDG-based, absolute in the future	End-value referred to in SDGs, e.g. infant mortality at 12 per 1000 lives	46
A2. SDG-based, relative to starting position	End-value referred to in SDGs, e.g. reduce by half the proportion of people living in poverty	6
B1. Other international agreement or shared aspirations, absolute in the future	End-value set by International Agreements, Good Practices or other Established Frameworks, e.g. reduce PM 2.5 pollution to less than 10 micrograms per cubic meter (WHO)	40
B2. Other international agreement or shared aspirations, relative to starting position	End-value set by International Agreements, Good Practices or other Established Frameworks, e.g. double the share of renewables in consumption (IRENA)	3
C. No explicit value; best historical performance considered	End-value set at the 90th Percentile of OECD countries in 2010	36

Source: OECD (2017c)

OECD Country Views on SDG Benchmarking

Well-positioned to capture data with robust education systems and consistent participation in international assessments, OECD countries have committed to pursuing the SDGs. Despite general commitment, the way in which these countries are benchmarking SDG 4 naturally vary, as well as their views on its indicators.

Estonia, for example, prefers that the EU set regional targets for the SDGs, stating that the current regional targets are too general. Avoiding political indicators is an additional point of emphasis. However, some potential competition has been noted between UNESCO and OECD in benchmarking efforts, with calls for better coordination between SDG 4, OECD 2030 and EU future targets (European Commission, 2017a). Even knowing what the role of benchmarks and indicators should be for greater impact and relevancy should be prioritized, according to some countries like Slovenia.

Germany in particular has been implementing a Sustainable Development Strategy since 2004, well before the SDGs adoption in 2015, though it has incorporated the 2030 Agenda into its revised strategy. The country's statistics office compiles the indicators to measure the goals and publishes its progress in Indicator Reports. Among its benchmarks are those recommended by the EU to measure SDG indicators (such as 4.1a), and those set by its federal government (e.g., indicator 4.2.b: all-day care provision for children 3- to 5-year-olds to be increased to 60 % by 2020 and 70% by

2030) (European Commission, 2017a). Complementing its Sustainable Development Strategy is one set by the federal government, which is set up along 12 dimensions with 46 indicators. Among its education dimensions are "chances for education for all" and "time for family and job."

In terms of measuring certain indicators and holding them to benchmarks, there are competing views on what needs to be used and what standards should be set. For example, while the European Commission's adult participation benchmark is relevant, some countries believe that participation rate is highly underestimated when using Labour Force Survey data, and suggest that Adult Education Survey data would provide more reliable estimates and influence benchmarks. Others consider using administrative register data more (instead of relying on survey data), since some countries' tertiary attainment benchmark rate is 5 percentage points lower compared to the Labour Force Survey data, as is the case in Sweden (European Commission, 2017a). Spain confirms the overall difficulties with SDG education indicators, emphasizing the importance of selecting indicators and benchmarks for their representativeness, their feasibility of calculation and temporality, being able to dispose of them in annual bases, if possible (European Commission, 2017a). The country also suggests the inclusion of qualitative data for a richer understanding of educational progress, a belief mirrored by several others.

B. Recommendations/Alternatives for Benchmarking SDG 4

Taking Advantage of International Assessments

A growing number of countries participate in cross-national learning assessments, whether regional or international, as well as hybrid assessments such as EGRA/EGMA and citizen-led assessments like ASER in India and Pakistan. While the hybrid and citizen-led assessments are not designed to compare countries' data against one another, some suggest that the rest of cross-national assessments can and should be leveraged in benchmarking SDG 4 to complement any national assessment a country currently implements (Birdsall et al., 2016). Because many of these assessments roughly correspond to measurement points under SDG 4, they provide a generally comparable global scale against which to measure the goal. Additionally helpful is the wide acceptance of not only these assessments' reliability and validity in capturing accurate data.

Some considerations must be made before taking this approach, including the target population and test construction. Treviño and Ordenes (2017) note that whether the assessment takes an age-based approach or a grade-based approach and whether it is curriculum content-based or competency-based have significant implications for the external validity and inferences made from comparison. There is also the criticism that multinational companies involved in these assessments have several and intricately interlinked interests in what they measure that must also be considered (Wulff, 2017).

Concern aside, for this strategy to succeed, developing reliable and valid items to cross-link the existing regional and international assessments would be a necessary step for both the assessments' governing bodies and countries' stakeholders. Countries, according to Birdsall et al. (2016), must have national assessments that incorporate items from regional and/or international assessments, as well as household surveys that include standardized learning modules to track learning outside the school environment. Some countries already draw upon items in these cross-national assessments to inform their national assessments: for example, Mexico plans to link its national assessment to PISA and has set presidential targets for 2012 and for 2030, while Brazil has

benchmarked every secondary school against PISA so that student performance is linked both to the national metric and to that of PISA's (NGA, 2008). While this eases measurement purposes, significant efforts to achieve linkages between the different assessments, as well as guaranteeing their reliability and validity, have been minimal. Research studies that attempt to link test scores using different assessments indicate the complexity of such a process (Sandefur, 2016; Treviño and Ordenes, 2017).

The participation in these assessments suggests a country's readiness in producing the data needed to not only improve student learning outcomes, but also measure progress towards SDG 4. The assumption in this recommendation, however, is that developing countries, in particular, have the capacity to participate in these cross-national assessments until they have both the ability to implement and availability of a national assessment. According to the Learning Assessment Capacity Index (LACI), the practice of using cross-national assessment items in national assessments is presently disjointed; for example, the majority of South American countries average between 4 and 5 assessments, whereas countries in Africa are between 0-2 (LACI, 2017). Investment from overseas development aid, which has typically lagged in spending on data and research compared to other sectors such as health, can address these capacity issues (Birdsall et all, 2016).

Interest in aligning cross-national assessments has been recently discussed among relevant actors in those assessments and UNESCO institutions. In 2017, several proposed to link regional assessments with TIMSS in 2019, in which two to three countries per region would participate in both the regional assessment and TIMSS in that year (Montoya & Hastedt, 2017). Results from both assessments would then be scaled to one another. The process, inspired by the Ring Comparison methodology, then allows remaining countries participating in each regional assessment to theoretically report on the TIMSS scale with the comparisons established by the original two or three countries. This will ideally allow for the establishment of minimum proficiency levels as benchmarks for monitoring learning (Montoya & Hastedt, 2017). Each region can also remain independent of other regions, whilst adopting the estimation methodologies that are best suited to its country characteristics and statistical capacities.

Treviño & Ordenes (2017) lay out four time-bound strategies for assessing SDG 4 that can build upon one another using data from cross-national assessments to ultimately create a Worldwide Proficiency Assessment. A summary of the strategies is listed in Table 4. The authors suggest that a mid-term specific instrument with a clear definition of a minimum level of competency for grades 2/3, end of primary, and end of secondary is the most technically appropriate approach to measuring indicator 4.1.1 due to the information available from the different cross-national assessments. They note, however, that no strategy is without its faults. The focus of the Worldwide Proficiency Assessment may seem too limited in capturing the complexity of SDG 4 in its measurement of only literacy and numeracy skills. Despite this, taking advantage of the current data that is available to measure SDG 4 and tackling the challenges of equating that data is necessary in order to understand the ways in which countries are progressing towards it.

Table 4. Summary of strategies for measuring SDG4

Strategy	Implications
Strategy 1: use of national assessments to measure SDG4 with adjustments using international assessments. To be implemented in the short run	High levels of external validity for measuring the minimum level of competency established in official curriculum. Low levels of international comparability
Strategy 2: equating among international and regional assessments. To be implemented in the medium run	Apparent low cost by using existing assessments. Entails performing one equating for each of the grades to be assessed in indicator 4.1.1 and defining new proficiency levels for each scale. Technically questionable from a psychometric and substantive point of view. Low levels of external validity for representing the national curriculum.
Strategy 3: equating between different international evaluations aiming at similar school grades. To be implemented in the medium or long run	Requires the definition of anchor items that can be shared across the different evaluations and the creation of a consortium of different assessment projects. Difficulties of comparison because of the differences in the domains assessed in the different assessments. Psychometrically and substantively more robust. Low levels of external validity for representing the national curriculum.
Strategy 4: creating a Worldwide Proficiency Assessment on Numeracy and Literacy. To be implemented in the long run.	Psychometrically and substantively robust. Politically difficult to convince countries to participate in this assessment. Requires the participation of technical institutions in the design, implementation, and analysis of test results. Low levels of external validity for representing the national curriculum.

Source: Trevino & Ordenes (2017)

Common Goals, Differentiated Targets

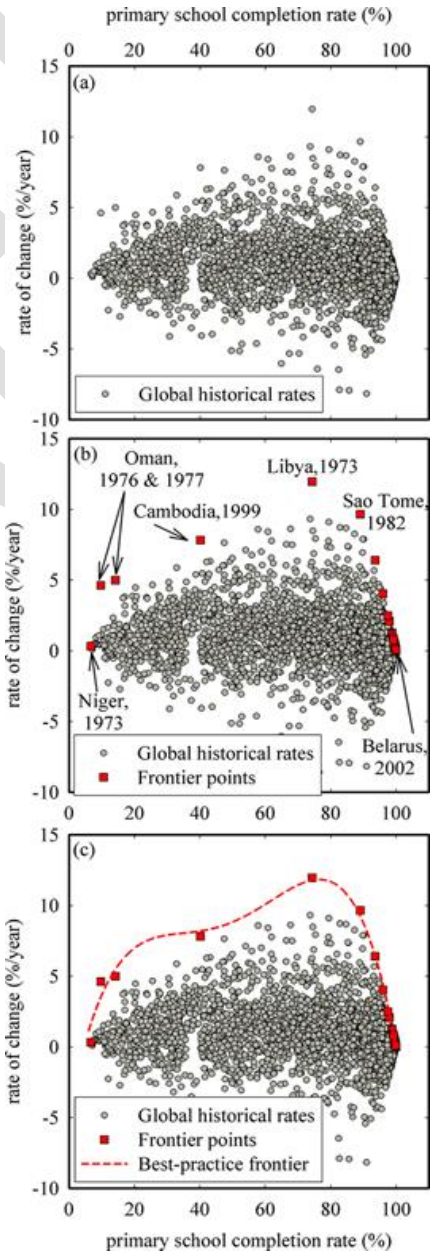
Questions regarding the breadth and depth of the SDGs, with concerns that they cover too much, or not enough, fuel the debate of the 2030 Agenda's effectiveness in creating positive change. Some view a one-size-fits-all benchmark as inappropriate in benchmarking the SDGs; rather, to truly make a difference in creating a better, sustainable world, the SDGs must be a set of common goals, but with differentiated targets (Martens, 2015; Adams & Judd, 2016).

Targets, in this opinion, should be divided into three types, per Martens (2015). The first type of target is one that creates global obligations to which all governments are bound, including rich country governments. The second is that which "does no harm," meaning that it forces countries to reduce negative external effects at the global level. Such an example is the reduction of per capita greenhouse gas emissions, as guided by SDG 13.2 (though specific levels are not mentioned). Finally, the third type of target directs attention to the responsibilities and duties of rich countries at the international level. These differentiations put much of the onus on those countries that not only make up the majority of resource consumption, but have the means with which to drive change (Martens, 2015). The last type has been turned into a global commitment, made in in Addis Ababa in the Finance for Sustainable Development Summit, and again in the 2030 Agenda and the Paris Climate Agreement, though particular details on accountability have yet to be determined (Sachs et al., 2016).

At the time of the SDGs' development, criticisms regarding the then-current targets were that they are not numerical and for those that are specific (such as the 0.7 target for ODA) they were not time-bound. For example, in measurements of inequality, one concern is that cautious targets may lead to inappropriate policy recommendations. For example, the target "by 2030, progressively achieve and sustain income growth of the bottom 40% of the population at a rate higher than the national average" solidifies the relationship between the reduction of inequality and steady economic growth, while ignoring the possible contributions of redistributing income and wealth (UN Working Group, 2014). Because income disparities have increased in recent years, the need to address the financial influences in inequality is crucial. One way this can be accomplished, according to Martens (2015), is utilizing the Palma ratio, which compares the income share of the top 10 percent to that of the bottom 40 percent of the population. Without focusing on the richest quintile as well as the poorest quintiles, achieving the goal on reducing inequality is as yet insufficient (Adam & Judd, 2016).

Holding each country to common goals but different targets also could influence the types of indicators used to measure them. Dill & Gebhart (2016) noted a redundancy and bias in leading international indices such as the Human Development Index and the Global Peace Index towards developed countries, meaning that the same countries tend to rank consistently highly in most of these indices. The contributing factor to this consistency, they argue, is the dependency of these indices on GDP, since many of the indicators tend to directly influence or are directly influenced by GDP. While developing countries are encouraged to achieve the SDGs,

Figure 1: Creation of frontier points. From Luh et al. (2016)



having indices that naturally favor European standards (which also happen to be the countries in which they are generally developed) may naturally disadvantage them (Clemens, 2004). Utilizing GDP as a standard against which to measure all countries against one another, according to Dill and Gebhart (2016), will not work in a post-2008 financial crisis world because of the unfair comparison of "real agricultural and commodities" economies with those countries "that can print their GDP growth by governmental bonds" (p. 8).

Using Frontier Analysis for Realistic Benchmarking

Drawing from the principles of data envelopment analysis, frontier analysis aims to identify benchmark rates using the rate of the historically best performing country among those at a similar level of coverage. For global targets in SDG 4, Luh et al. (2016) suggest that frontier analysis can be used to benchmark country performance. Using the indicator "primary school completion rate", Luh et al.(2016) first identify frontier points based on which countries within the same level of coverage have the highest rates of changes and completion rates at a given year (see Figure 1). Performance indices are then created with the following formula:

$$index = \frac{country\ rate - minimum\ rate}{maximum\ rate - minimum\ rate},$$

By calculating the maximum possible rate (also referred to as the benchmark rate) to create performance indices, countries can theoretically be fairly compared to one another. The major assumption of this method, however, is that countries at similar levels of coverage will be able to produce the same, if not similar, achievements towards the benchmark.

According to Luh et al. (2016), the main advantage of using frontier analysis is its ability to use both status and rate of change in such a way that eliminates the biased starting point of each country. In addition, frontier analysis allows for the identification of "unfulfilled potential a country has to most effectively use its resources to achieve the greatest possible progress" (Luh et al, 2016).

Relatedly, Fukuda-Parr and Randoph (2014) proffer the Achievement Possibilities Frontier (APF) as an approach to benchmark SDGs using per capita income levels. Country-level data for a specific indicator – again, using primary completion rate as an example – is first plotted against per capita income for a specific time. Then, through econometric techniques such as a curve fitting routine, a frontier, is created to benchmark achievable targets at any given income level within a country. For example, if the data shows that at a per capita income level of $7800, the best achievement is 100% of students completing primary school, then countries with per capita income levels greater than $7800 would be held to a benchmark of 100%. If a country with a $2000 per capita income level is best able to reach 70%, then countries within that level would be held to said benchmark, and so on. Thus, by creating a universal standard within a country's economic means, the APF approach can set feasible targets for future achievement in the SDGs.

Frontier analysis does not necessarily account for shocks to an individual country or other circumstances that may lead to under- or over-performance. Another concern is that this method is randomly determined, having a distribution or pattern that may be analyzed statistically but may not

be predicted precisely. Furthermore, monitoring other indicators within SDG 4 that would be considered more difficult to measure, such as gender equality in education beyond parity, may not be able to be captured with this method. As such, Luh et al. (2016) suggest this method as a complement to monitoring the SDGs, especially considering their previous successful work using the frontier analysis in water and sanitation in developing benchmarks, and has been suggested as a method to assess equity, a major theme in the SDGs (Luh et al., 2016).

C. Concerns and Challenges

An Ambitious Agenda

SDG 4 and its associated targets present an ambitious agenda with a focus on quality learning and equity in education in addition to the more traditional indicators of access and participation. As the OECD report mentions, "it challenges every single country in the world to improve its education system and marks a significant departure from previous global education goals and targets, such as the Millennium Development Goals (MDGs) and Education for All (EFA), that were not universal and focused more on access and participation" (OECD, 2017a). Therefore, there lies ahead a monumental challenge with respect to achieving targets that measure learning outcomes and equity (OECD, 2017a).

Countries' Capacities

The scope of the monitoring agenda is wide, and the issues involved are often complex. The types of data that contribute to understanding how learning occurs vary widely, from political dynamics, to education system characters, inputs (e.g., textbooks and teachers), and service delivery indicators, among many others (Birdsall et al., 2016). Another concern is the financial capacity for countries to collect the data necessary to monitor progress on SDG 4. Therefore, from one perspective, the Education 2030 monitoring agenda, when presented individually by target, is daunting in its breadth. Responding to each target would pose significant challenges to education ministries and national statistical agencies, not only in poor countries but also in rich ones (UNESCO, 2016).

Political Economy

Because definitions matter, one significant challenge in establishing benchmarks ends up being not only a validity issue, but also a political one. The conversation surrounding development, and by extension, the global goals, is that the phenomenon has seen as technical rather than political. Bishop (2016) claims that, as a result, global goals only represent symptoms of development rather than development itself, and that policy aiming to achieve those goals will lead to incremental and not transformative growth.

The political economy, therefore, is intimately related to the process of setting benchmarks for SDG 4. In the past decade, the BRICS – Brazil, Russia, India, China and South Africa – and even the MINT countries – Mexico, Indonesia, Nigeria and Turkey – have all dramatically increased their collective share of the global GDP, without signs of slowing down (Bishop, 2016). These countries not only have fundamentally different political economies and distinctive patterns of state-society relations and political competition, but they are also exploring a range of strategies to achieve their development ends. Relying on different assumptions and enabling different implications, these strategies offers the international community the "opportunity to rethink what development might be and why it matters" (Bishop, 2016: 10). As Ghosh (2015: 321) argues, development is not really

"about simply reducing deprivation", as envisaged in universalist MDGs and SDGs, which tend to strip history and context from societal challenges, it is more about "transformation—structural, institutional and normative—in ways that add to a country's wealth-creating potential, ensuring the gains are widely shared and extending the possibilities for future generations."

Consensus on Acceptable Benchmarks

Reaching an agreement on what constitutes an acceptable minimum benchmark against which to monitor education development requires global discussion and conclusions, which according to Treviño & Ordenes (2017) face three challenges. The first challenge is how well national curricula and their various learning objectives match that of the global definition of minimum level of competency. The second is the how the results of using such a definition will influence how poorly-performing countries are identified, and naturally highlight the potential biases of a global definition. The third challenge centers on guaranteeing the external validity of the assessment utilized to record proficiency, as issues such as those brought up by the first challenge may lead nations to question its results. Eliminating hierarchical categories in favor of a continuum promotes the truthful view that development is contextually mediated and rooted, yet still requires a transformative process of change. Improving living standards is important, but this will not happen in an even or truly fundamental way by blindly pursuing the SDGs as an end in themselves (Bishop, 2016). As such, what ends up being measured is crucial.

Other Challenges and Concerns

In this view, the proposed monitoring agenda is underwhelming. Considering ways in which to monitor the indicators with the paucity of data on important components makes the monitoring agenda seem as though it is not doing enough. Compounding this further is the lack of information on several of these aspects covered by SDG 4, for example the impact of in-service training on actual classroom behavior (UNICEF, 2017). Concerns about accurate representation and comparability with the current data too must also be recognized; Sandefur (2016) found that students in African countries taking the SACMEQ test were two full grades behind most of those taking the TIMSS assessment in grade eight. Given the severity of sustainable development challenges, the monitoring agenda barely scratches the surface of the fundamental questions facing education and lifelong learning. How do education systems help learners of all ages acquire relevant knowledge, practice critical thinking, manage uncertainty, act responsibly regarding the environmental crisis, understand their shared humanity and behave as global citizens? The monitoring framework does not go far enough to answer these questions (UNESCO, 2016).

Another concern is that the SDGs do not report on structural concerns and the responsibilities of duty-bearers (King, 2015; Adams & Judd, 2016; Wulff, 2017). In the case of education, governments can report on enrollment figures and learning outcomes, yet are not obligated at all to disclose who provides such education and how much it costs citizens. They are also free to report on whichever targets they choose. Such absence of rights-based indicators within education may fail to fully and accurately represent the quality and equitable aspects of the goal, as well as negatively influence progress on SDGs 5 (gender equality), 8 (decent work), and 10 (inequality) (Clemens, 2004; Wulff, 2017).

Yet, overall, important steps have been taken. The agenda is instilling a new sense of purpose in education monitoring activities, compared with the emphasis of past decades on access and

participation. While there are concern that the SDG 4 targets and indicators may eventually direct and narrow efforts by governments and IOs alike in order to achieve actionable results (Sprunt et al., 2017; Wulff, 2017), the expanded approach to reviewing progress should be celebrated and safeguarded. It offers a starting point for advancing the sustainable development agenda, with education at its heart. The challenge is for governments and the international community to take concrete steps to achieve the new education targets while acting swiftly and purposefully to enable their monitoring, based on agreed indicators, even those with identified weaknesses.

The establishment of a permanent group for technical cooperation, representing a large number of countries, is a considerable advance in the international dialogue on education monitoring, and fills a notable gap experienced during the Education for All period. At least two challenges lie ahead. First, countries must be assured an opportunity to contribute to discussions in an informed and meaningful way. Their active role in the TCG is critical. Second, a mechanism is needed for future decision making within the TCG, to help reach consensus and strengthen the group's legitimacy (UNESCO, 2016).

Table 5: Summary of Views/Alternatives on SDG 4 Benchmarking

Views/ Alternatives	Advantages	Disadvantages	Notes
Incheon Declaration	• Aims for an inclusive process, with full transparency and accountability, and engaging all partners. • Country ownership and common understanding. • Intermediate benchmarks could be set for each target to serve as quantitative goalposts for review of global progress vis-à-vis the longer term goals. • Such benchmarks to be built on existing reporting mechanisms, as appropriate.	• Involvement of all the partners through an inclusive process, with full transparency and accountability is challenging and time-consuming process. • Governments have differing education priorities, institutional capacities, and availability of resources.	• There is a need to establish appropriate intermediate benchmarks (e.g., 2020, 2025).
OECD	• The indicators selected have face validity, discriminatory power, broad availability, and high statistical quality. • Ability to cover about 100% of SDG 4	• Data cannot cover 57% of the overall SDG targets to be evaluated. • Significant statistical work is needed to fill some important gaps. • Low-performers (and minimum proficiency levels) may not be accurately captured or represented by the data.	• Built on the UN global indicator framework. • Measures countries' relative distances from SDG targets, with available data. • Identified 131 indicators covering 98 targets spanning all 17 Goals.

Views/ Alternatives	Advantages	Disadvantages	Notes
Using of International Assessments	• Many of these assessments roughly correspond to measurement points under SDG 4 and are already use in its monitoring. • Provide a generally comparable global scale against which to measure the goal. • Wide acceptance of these assessments' reliability and validity in capturing accurate data. • Some countries already draw upon items in these cross-national assessments to inform their national assessments. • Ideally allows for the establishment of minimum proficiency levels as benchmarks for monitoring learning. • Each region can remain independent of other regions, whilst adopting the estimation methodologies that are best suited to its country characteristics and statistical capacities.	• Multinational companies involved in these assessments may influence what is measured. • Assumes that countries have national assessments that incorporate items from regional and/or international assessments, as well as household surveys that include standardized learning modules to track learning outside the school environment. • Achieving linkages between the different assessments, as well as guaranteeing their reliability and validity is a complex and challenging process. • Developing countries, in particular, might not have the capacity to participate in these cross-national assessments.	• These assessments can and should be leveraged in benchmarking SDG 4 to complement any national assessment a country currently implements.
Common Goals, Differentiated Targets	• Places responsibility on all countries, not just developing ones, to achieve the SDGs. • Addresses the financial influences in inequality.	• Some of the rich countries might not be willing to accept this recommendation. • Defining or benchmarking differentiated target is complicated. • Comparisons using differentiated targets might not be so reliable. • Current indices to measure SDG progress may be biased towards developed countries. • Structural concerns and the responsibilities of duty-bearers are not yet addressed by the SDGs.	• The proposition is that SDGs must be a set of common goals, but with differentiated targets. • National percentages exist for many of the SDG 4 targets.

Views/ Alternatives	Advantages	Disadvantages	Notes
Frontier Analysis	• By creating performance-based indices, countries can theoretically be fairly compared to one another. • Ability to use both status and rate of change in such a way that eliminates the biased starting point of each country. • Allows for the identification of unfulfilled potential a country has to most effectively use its resources to achieve the greatest possible progress. • Has been used successfully in setting water and sanitation benchmarks.	• Countries at similar levels of coverage might not be able to produce the same, if not similar, achievements towards the benchmark. • Does not necessarily account for shocks to an individual country or other circumstances that may lead to under- or over-performance. • Method is randomly determined, having a distribution or pattern that may be analyzed statistically but may not be predicted precisely. • Monitoring other indicators within SDG 4 difficult to measure (e.g., gender equality in education beyond parity), may not be able to be captured with this method.	• Suggested to be used as a complement to monitoring the SDGs, especially considering their previous successful work using the frontier analysis in water and sanitation in developing benchmarks. • Suggested as a method to assess equity, a major theme in the SDGs.

II. Existing Benchmark Initiatives

While the new focus on monitoring learning outcomes and the quality of education is a welcome shift in SDG 4, a consensus on what should be measured has yet to be reached (UNICEF, 2017). This, however, has not stopped several international, regional and national efforts to benchmark the global education goal and begin progress towards it. This section identifies some of these initiatives both in terms of benchmarking efforts of SDG 4 as well as standardization and other benchmarking initiatives pertaining to the global goals. A summary of these initiatives along with their advantages and disadvantages is included in Table 7 at the end of this section.

The International Standard Classification of Education

The International Standard Classification of Education (ISCED) is the standard framework used to categorize and report cross-nationally comparable education statistics that belongs to the United Nations International Family of Economic and Social Classifications. Initially developed by UNESCO in the 1970s, and first revised in 1997, the ISCED classification serves as an instrument to compile and present education statistics both nationally and internationally. ISCED is applied in statistics worldwide with the purpose of assembling, compiling and analyzing cross-nationally comparable data. ISCED is the reference classification for organizing education programs and related qualifications by education levels and fields. Its 2011 classification was adopted by the UNESCO General Conference at its 36th session in November 2011. The framework is occasionally updated in order to better capture new developments in education systems worldwide (UIS, 2012).

ISCED is designed to serve as a framework to classify educational activities as defined in programs and the resulting qualifications into internationally agreed categories. The basic concepts and definitions of ISCED are therefore intended to be internationally valid and comprehensive of the full range of education systems. ISCED classifies education programs by their content using two main

cross-classification variables: (a) levels of education, and (b) fields of education. ISCED 2011 presents a revision of the ISCED 1997 levels of education classification. It also introduces a related classification of educational attainment levels based on recognized educational qualifications (UIS, 2012).

Information compiled according to ISCED can be used for assembling statistics on many different aspects of education of interest to policymakers and other users of international education statistics. These aspects include enrolment and attendance, human or financial resources invested in education, and the educational attainment of the population. The application of ISCED facilitates the transformation of detailed national education statistics on participants, providers and sponsors of education, compiled on the basis of national concepts and definitions, into aggregate categories that can be compared and interpreted internationally (UIS, 2012).

ISCED 2011 rests on three components: i) internationally agreed concepts and definitions; ii) the classification systems; and iii) ISCED mappings of education programs and related qualifications in countries worldwide. The basic units of classification in ISCED are the national (and sub-national) education program and the related recognized educational qualification (UIS, 2012).

European Union/OUE

The UNESCO-UIS/OECD/EUROSTAT (UOE) data collection is administered jointly by the United Nations Educational, Scientific, and Cultural Organization Institute for Statistics (UNESCO-UIS), the Organisation for Economic Co-operation and Development (OECD), and the Statistical Office of the European Union (EUROSTAT). The objective of the joint UOE data collection on education statistics is to provide internationally comparable data (mostly at national level, with some insights at the subnational level) on key aspects of formal education systems, specifically on the participation and completion of education programs, as well as the cost and type of resources dedicated to education. Countries participating in the UOE data collection co-operate to gather the information, to develop and apply common definitions and criteria for the quality control and verification of the data. This participation is crucial to the EU's open method of coordination (OMC) for benchmarking procedures, though the method has been contested by the literature for its ineffectiveness due to poor design and overlooking the need for different types of benchmarking (Groenendijk, 2009). In addition to the metadata asked for in the different questionnaires, EU, EFTA and EU candidate countries provide standard data quality reports as requested by Commission Regulation (EU) No 912/2013 (UOE, 2016a).

National education indicator frameworks for the EU have been developed since the start of the century; the Education and Training 2020 (ET 2020) is the strategic framework for the EU provides common strategic objectives to help Member States further develop their educational and training systems. The framework takes into consideration the whole spectrum of education and training systems from a lifelong learning perspective, covering all levels and contexts (including non-formal and informal learning). The original benchmarks were chosen based on their comparable data and the differing situations in individual Member States (European Union, 2017b). Though there are more indicators in the framework, the ET 2020 defines the following seven education benchmarks to be achieved by countries:

- At least 95% of children (from 4 to compulsory school age) should participate in early childhood;

- Fewer than 15% of 15-year-olds should be under-skilled in reading, mathematics and science;

- The rate of early leavers from education and training aged 18-24 should be below 10%;

- At least 40% of people aged 30-34 should have completed some form of higher education;

- At least 15% of adults should participate in lifelong learning;

- At least 20% of higher education graduates and 6% of 18-34 year-olds with an initial vocational qualification should have spent some time studying or training abroad;

- The share of employed graduates (aged 20-34 with at least upper secondary education attainment and having left education 1-3 years ago) should be at least 82%.

Though the benchmarks are used and disseminated widely at national levels, there are no mandates requiring countries to take up all ET 2020 benchmarks. Many countries, therefore, have their own national standards against which to measure their learners, as is the case with Denmark (e.g., one benchmark is to "increase the well-being of pupils") (European Commission, 2017a). The indicators and benchmarks often are modified by countries to meet their specific context in their national plans, whether by changing the definition or using a different data source, or by adopting targets that align with their own ambitions for development (European Commission, 2017b). This flexibility can be seen as one of the strengths of the ET 2020, as long as the benchmarks are aligned with national frameworks. At the same time, as is the case with its OMC method in general, the EU seems to suffer from a mix of objectives for its benchmarking activities (Groenendijk, 2009).

Future discussions for post-2020 ET benchmarks feature indicators and topics that, while not yet covered by the ET 2020, already are incorporated into both national/EU indicators or policy directions and the SDG 4 targets (European Commission, 2017b). At the same time, emphasis has been placed on having a limited number of benchmarks to increase their impact. Some of the strategic objectives in ET 2020 involve concepts that are hard to measure (e.g., 'creativity', 'innovation', 'quality') and there are no metrics that are readily available to robustly understand achievements, which explains why some indicators did not receive specific benchmarks.

This all highlights the tension in prioritizing certain indicators over others, and spurs debate as to whether benchmarks should be set for all indicators in the framework. Several arguments can be made against such a task. First, the percentages set in the ET 2020 benchmarks may not necessarily work for certain indicators based on country context. For example, taking into account the different migration patterns across Europe, establishing a percentage for foreign-born early school leavers may mean significant changes in or evaluations of France's education system than it does for Poland's. Such considerations would certainly lead to intense debates about establishing a common benchmark. As noted by Groenendijk (2009), as a diversity-accommodating, bottom-up policy-learning device, OMC-benchmarking does not work due to limited member state participation and poor vertical flow of information and ideas. To guarantee consensus may be an overreaching expectation.

Additionally, education systems may not have the appropriate measures or structure to measure certain indicators for benchmarking purposes. In higher education, ET 2020 sets an indicator on recognition of informal and non-formal learning as an entry to higher education, yet for the majority

of EU countries, these types of learning are not recognized (European Commission/EACEA/Eurydice 2016). Because some priorities highlighted in ET 2020, as well as SDG 4, have yet to have thorough measurement standards, establishing a benchmark for other indicators may not yet be appropriate. This is not to suggest that because the data does not yet exist that there should be no efforts to measure it; rather, before a benchmark is set, it is crucial to better understand the factors involved in its measurement. Because the EU is seen as an "important mirror and benchmark for countries across the world," its decisions will have widespread influence (Kamens, 2013: 124).

OECD

During the 1980s, the increasing demand for information on education and the need for improved knowledge about the functioning of education led authorities in the OECD member countries to consider new ways of comparing their education systems. They reached agreement on the feasibility and utility of developing an international set of indicators that would present, in statistical form, key features of their education systems (OECD, 2017b).

The OECD's Centre for Educational Research and Innovation (CERI) responded to this demand for comparative information by initiating the OECD's Indicators of Education Systems (INES) Programme. The program developed a provisional framework for organizing the indicators, proposing a set of indicators and the methodologies for measuring them. This framework has been considerably developed since then and is summarized below.

The first set of indicators was published in Education at a Glance in 1992 and drew mainly on existing data sources. The work to produce the first Education at a Glance exposed weaknesses both in the underlying statistical classification (the ISCED) and in the data collections themselves. Since then, much work has been put into revising ISCED and improving the methods and instruments for the international data collection on education (OECD, 2017b).

The OECD education indicators are the product of an ongoing process of conceptual development and data collection with the key objective of linking a broad range of policy needs with the best available international data. Benchmarking in the OECD follows a rigorous structure plan, which involves planning and defining the area of study, collecting, structuring and evaluating data, and reviewing and revaluating policy domains to identify effective approaches (Groenendijk, 2009). In each area of work, the following considerations have, traditionally, guided the indicator activities (OECD, 2017b): (a) emphasizing those education issues where the international comparative perspective can offer significant added value over and above what can be achieved through national analysis and evaluation; (b) seeking to strike an appropriate balance between focusing new developments on areas where the feasibility of data development is promising, and not neglecting important areas where substantial investment in conceptual and empirical work is needed to further the policy debate, and (c) continually reviewing the work to ensure that the outcomes are cross-nationally valid and reliable (OECD, 2017b).

The indicator program places increasing emphasis on integrating its work through the perspective of lifelong learning, with the aim of progressing from a model of education built around institutions to one that looks more broadly at the extent and benefits of learning throughout life. In addition, various activities within the program are seeking to better reflect equity-related issues, through assessing differences and inequalities among individuals and groups of individuals (OECD, 2017b).

The OECD's education indicators claim to address the issue of measuring the current state of education internationally. They provide information on the human and financial resources invested in education; access to education, progression, completion and transitions from education to work; the learning environment and the organization of schools; the quality of learning outcomes; and the economic and social returns to learning. The education indicators are organized thematically and each is accompanied by relevant background information. The indicators are presented within an organizing framework with the following features:

• Distinction between the actors in education systems: individual learners, instructional settings and learning environments, educational service providers, and the education system as a whole

• Grouping of the indicators according to whether they are measures of learning outcomes for individuals and countries, policy levers or circumstances that shape these outcomes, or antecedents or constraints that set policy choices into context

• Identification of the policy issues to which the indicators relate, with three major categories distinguishing between the quality of educational outcomes and educational provision, issues of equity in educational outcomes and educational opportunities, and the adequacy and effectiveness of resource management (OECD, 2017b). Table 6 below shows the first two dimensions of this framework.

Table 6. OECD Education Indicator Matrix

Education indicator matrix

	(1) Education and learning outputs and outcomes	(2) Policy levers and contexts shaping educational outcomes	(3) Antecedents or constraints that contextualise policy
(A) Micro-level: Individual participants in education and learning	(1.A) The quality and distribution of individual educational outcomes	(2.A) Individual attitudes, engagement, and behaviour	(3.A) Background characteristics of individual learners
(B) Meso-level: Instructional settings	(1.B) The quality of instructional delivery	(2.B) Pedagogy and learning practices and classroom climate	(3.B) Students' learning conditions and teachers' working conditions
(C) Meso-level: Providers of educational services	(1.C) The output of educational institutions and institutional performance	(2.C) School environment and organisation	(3.C) Characteristics of service providers and their communities
(D) Macro-level: The education system as a whole	(1.D) The overall performance of the education system	(2.D) System-wide institutional settings, resource allocations, and policies	(3.D) The national education, social, economic and demographic context

Source: OECD (2017b)

Advantageous Homogeneity

Having a clear objective for its benchmarking activities allows the OECD to avoid many pitfalls associated with benchmarking, it is important to note the distinct characteristics of the organization that allows it to set benchmarks with relative ease. Participation in the OECD depends largely on mutual trust between member states and shared confidence in the process, with the goal to learn from one another. Indeed, the organization relies on good arguments and a common value system to influence national policy makers (Groenendijk, 2009). Because of the general economic prosperity and homogeneity of member states, as well as its function as an ideational agency and being separated from politics as much as it can, there is little competition among participating states. All

these factors serve to smooth the process of establishing benchmarks for a group of countries that are more alike than they are different, a far departure from the demands that a global education goal requires.

Unilateral Threshold Decisions

International organizations (IOs) like the OECD exercise power by organizing three types of supposed apolitical and technical actions. First, they classify by stratifying countries according to their level of performance in international evaluations such as PISA and, according to their results, put governments under great pressure to introduce education reforms to achieve better scores. Secondly, they fix meanings in the social world by, for instance, defining what educational development means. This is something that IOs can do explicitly, but also indirectly in the form of indicators and benchmarks. Thirdly, IOs articulate and disseminate new norms, principles and beliefs by, for instance, spreading what they consider 'good' or 'best' practices' in educational development.

The OECD, in response to the new 2030 Agenda, undertook all three steps. Confident in its measurement tools, the OECD explicitly states that measuring the SDGs can be achieved with data it already offers through instruments like PISA, PIAAC, and TALIS. In its 2016 edition of *Education at a Glance,* the organization includes a table of its measurement tools corresponding to each SDG 4 indicators it can monitor. More problematically, the report also charts OECD countries' current performance against a new quantitative benchmark for each of the SDG 4 targets (OECD, 2016a). The benchmark is seemingly calculated as the unweighted mean of available data values for each target, but still requires "more sophisticated approaches … to reflect the various facets of the targets and global indicators" (OECD, 2016a: 14). The benchmarks across the SDG targets in the report, however, closely mirror OECD and EU22 averages, thus establishing them as the "best" countries can aspire to.

Such a unilateral decision, however, tends to create a situation where countries are benchmarking more as a compulsion rather than their own choice. While representing a more homogenous group of countries, there is no doubt of the OECD influence over not only the international assessment field, but also that of education development, as non-OECD countries – as well as other institutions – look towards this group as a yardstick against which to measure their own progress in many aspects of their education systems or benchmarking efforts (Groenendijk, 2009; Kamens, 2013). As Kamens (2013) writes, the OECD's transformation of assessment and benchmarking into "a badge of good citizenship," fuels a competition between OECD countries and emerging ones, despite the OECD benchmarking approach being more cooperative in nature within OECD countries. Countries, therefore, are now indirectly held to these standards, when mutual agreement on those standards was not reached among countries/institutions.

Regional Assessments

As mentioned earlier, during the past two decades, the quantity, frequency, and systematization of cross-national student learning assessments around the world has increased, with broad implications for capturing the data required of SDG 4. Some of the major regional assessments – SACMEQ, LLECE and its iterations, and PASEC – and their benchmarking initiatives will be discussed to highlight how other cross-national assessment benchmark their learners.

All three assessments use a curriculum-focused, grade-based sampling design and similar methodologies. Before each assessment, LLECE member countries review their curricula and identify common content and cultural roots; later iterations also included UNESCO's 'skills for life' approach in their design (ACER, 2014). SACMEQ assesses Grade 6 students in reading and mathematical skills and puts students into eight different proficiency levels for both subjects (UIS, 2017f). The assessment uses Rasch item-response theory to establish the difficulty level for each test item; similar processes were employed in the LLECE and PASEC assessment. The LLECE assessment sets an expected percentage of students for four levels, and the distribution of students across these levels is compared between countries (ACER, 2014). PASEC was designed to analyze students in Grade 2 and 6 in Francophone Africa, and divides its students into 4 levels for reading and three for math.

In terms of minimum proficiency, all these assessments provide some form of a standard against which to benchmark their sampled learners based on what they can do at certain competency levels. For SACMEQ, of the eight proficiency levels, the third one considers students as reaching the basic competencies in reading and mathematics (UIS, 2017f). For example, in reading, basic proficiency is described as being able to "interpret meaning (by matching words and phrases, completing a sentence, or matching adjacent words) in a short and simple text by reading on or reading back" (Hungi et al., 2010). The TERCE assessment does not overtly suggest any minimum level of proficiency, though UIS views its Level II as a minimum one based on analysis of competencies acquired by pupils at that level (UIS, 2017f). This "sufficient" performance threshold of 40% of correct responses on the exam is considered the minimum level in the PASEC assessment. The benchmark is defined on the basis of the concepts assessed in the test and relies on the curricula goals in both reading and mathematics for Grade 2 and 6 (PASEC, 2015). While concerns about reliability in comparing the minimum level benchmarks set by these assessments exist, it is useful to understand how these benchmarks were developed and in what ways can they be leveraged in measuring SDG 4.

Qualifications Frameworks

The development of National and Regional Qualifications Framework (NQFs and RQFs) has been a major international trend in reforming national education and training systems since the 1990s, though in recent years their development has become more widespread. NQFs are now being implemented or developed in over 150 countries as of 2015 (UIL, 2015a), while eight RQFs support cross-border mobility of learners and workers and acting as a means for fair and transparent recognition of qualifications. QFs make significant impact in their ability to introduce different levels of standards which describe the characteristics and context of learning at is expected at each level. As they aim to improve not only the quality of worker qualifications but also their relevancy in the modern workplace because of their push to reform technical and vocational education and training (TVET). For example, Mauritius has established over 4,400 standards across 23 sectors of its economy since developing its NQF in 2004 (UIL, 2015b). RQFs are particularly helpful in their ability to set common standards for competences, such as that developed by the Association of South East Asian Nations (ASEAN). In many European countries, new learning outcomes and setting future targets in a systematic way have been inspired by NQFs.

Within NQFs are set standards against which to benchmark the development of TVET skills, which contribute to monitoring an important new aspect of SDG 4. For example, reforms in Bangladesh to

the TVET system developed new qualifications that not only ensure their quality, but helped establish a benchmark for comparison with international standards, while Hong Kong's Education Bureau utilizes its own NQF as a benchmark for quality in continuing education (UIL, 2015b). The shift towards defining qualifications by learning outcomes that fit NQF descriptors provides education authorities the ability to develop more relevant curricula for learners to develop the skills to succeed. Additionally, having a standard set of qualifications that respond to and benchmark increasingly globalized industries can catalyze a more systematic approach to skills development and measurement in education. Chakroun (2017) develops this relationship further, noting the "mutually reinforcing" nature between the Education 2030 Agenda and QFs (p. 11). Together, he states, SDG 4 and NQFs can improve learning assessments by focusing attention on developing appropriate standards for assessments, especially considering the latter are being created or reformed with identifying relevant skills for the workforce specifically in mind.

As QFs are referenced to one another, understanding the methodology of aligning different country levels to one another, as well as determining skill progression across different domain, will play key roles in establishing a proper global and independent reference metric with which to compare learning internationally. Chakroun and Daelman (2015) have noted the increase in benchmarking between QFs, typically between the national and regional frameworks, which suggests a general trend towards developing common tools to recognize learning. Thus, in 2012, in response to an increasingly global workforce and job landscape, UNESCO began developing an initiative to create a set of world reference levels (WRLs) aimed at representing generalizable indicators of levels of learning and serving as a global metric with which countries can compare different types of learning (Chakroun & Daelman, 2015).

Chakroun and Ananiado (2017) propose a three-dimensional structure based on analysis of QFs for the WRLs with the following dimensions: 1) a small number of broad stages; 2) a range of key factors and markers, in neutral language, that are common to most QFs; 3) definitions and explanations of said factors and markers; 4) a series of lists of these factors and markers that set the stages, with the aim to link WRL terms to those used in QFs. Each stage is then described with statements based on 11 factors, including carrying out activities, using and extending knowledge, and applying values (Chakroun & Ananiado, 2017). Though these are currently under discussion, they offer a promising start to conceptualizing WRLs.

Global Alliance to Monitor Learning (GAML)

Concerned with the plethora of regional and international learning assessments with little harmonization between their outcomes and what they inform relevant stakeholders, the Global Alliance to Monitor Learning (GAML) was conceptualized by UIS in early 2017 with two basic objectives: 1) to support national strategies for learning assessment, and 2) to ensure international reporting on the SDGs by all UN member states (UIS, 2017b). The GAML aims to bring together national education authorities, assessment agencies, citizen-led initiatives and the international education community, including donors, to ensure that countries have the high-quality data needed to improve the learning outcomes of all and to track progress globally. Underpinning the GAML is the belief that better data production and better use of that information at both the national and global level will not only streamline the performance of education systems, but also lead to improved learning outcomes of children and youth (UIS, 2017c).

As per its Result Framework (UIS, 2017b), the GAML functions in two tracks. At the national level, the GAML works with partners to develop tools, standards and guidelines to help countries who do not have national learning assessments to develop one, and for those that do, to improve their efficiency and efficacy in utilizing that data. At the global level, the GAML seeks to establish a common framework and data validation process for quality global reporting, and will work with the

Global Partnership for Education (GPE) and the Assessment for Learning (A4L) initiative. Under the auspices of UIS, the GAML intends to take the lead in providing these capacity-building services and coordinating data to fit both the usefulness at the national level and the comparability at the global level to achieve SDG 4 (UIS, 2017a).

The GAML, along with the ACER Centre for Global Education Monitoring, are exploring the development of the UIS Reporting Scales. Their goal is to support the use of existing national and cross-national assessments to facilitate measurement and reporting of learning outcomes (ACER, 2017). Each reporting scale starts by mapping cross-national and national curricula and assessment frameworks, to support international consistency in reporting among the 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.

The discussion paper prepared by ACER for the 4th GAML meeting provides some useful recommendations. Although they are in the context of benchmarking of the UIS Reporting Scales, they are relevant to the global benchmarking efforts as well. Their recommendations are as follows (ACER, 2017):

1. Use ISCED to in providing a cross-nationally standardized way of referring to the measurement points in Indicator 4.1.1. However, the in-precision in terms of years of schooling and applicability to out of school cohorts will need to be considered.

2. Countries' specifications for the target grades that correspond to the measurement points in indicator 4.1.1 will need to be adjudicated against an agreed set of criteria.

3. Adopt more precise interpretations (i.e., than the current 'Grade 2/3', 'the end of primary' and 'end of lower secondary') for the target groups and consider the implications for an out-of-school equivalent.

4. Adopt a single descriptive definition of the standard (i.e., minimal proficiency) for all three grade levels.

5. Though benchmarks should be content-referenced, their establishment should be informed by normative data where such data exist.

6. Agreement will need to be reached about the interpretation of the expression 'minimum proficiency' for each measurement point for each domain. A process for achieving this agreement is required.

7. The establishment of the benchmarks on the UIS scales will need to be: (a) informed by curricula from a variety of countries; (b) An iterative process; (c) Consistent with existing national and international standards.

8. Consideration will need to be given to whether the benchmarks should be points or ranges on the UIS scales.

9. The approach adopted to set the benchmarks should be primarily test-centered but include more than one method and, where possible, draw on performance data to support judges' decisions.

10. The establishment of the benchmarks will require the establishment of a panel of experts. The panel should be: (a) Selected from national nominees; (b) Have a high level of expertise in education in the relevant learning domain; (c) Large; (d) Well resourced.

11. Expert panels will need to be supported to develop and consolidate clear schema of what minimal proficiency in the domain looks like at the relevant measurement points. Support comes through providing them with high-quality training and clear, unambiguous definitions of key terms.

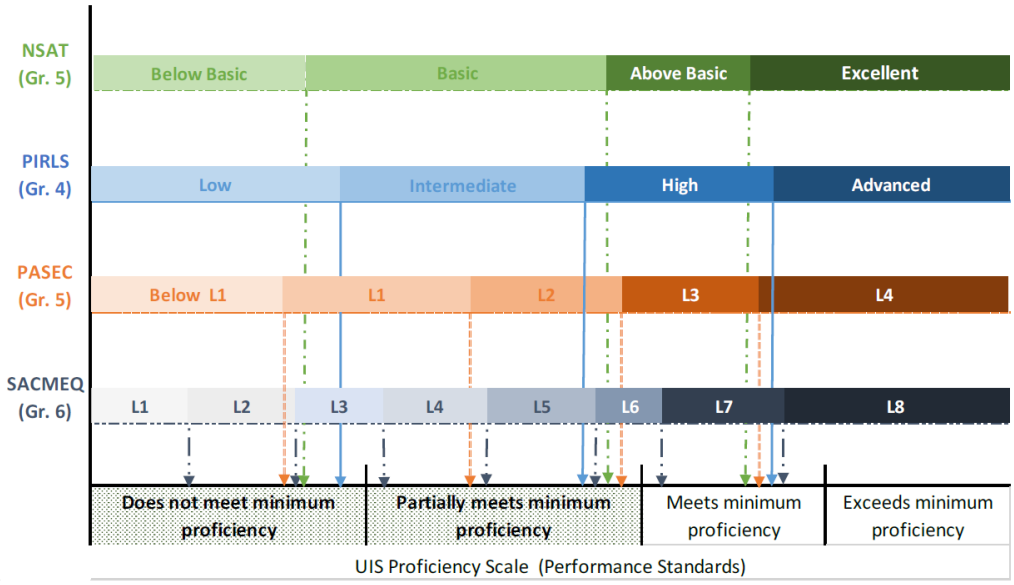
12. The benchmarks set by the expert panels should be submitted to broader stakeholder consultation before finalization

UIS and GAML are working with Management Systems International to address the immediate need of defining "minimum proficiency levels" for reading and mathematics for reporting the indicator 4.1.1.

These agencies are also working to produce a reporting metric and a mechanism for linking existing assessments and their performance levels to this metric (UIS, 2017e). Their recent discussion paper presents the following three steps involved in constructing a "UIS proficiency scale": (1) Defining content standards; (2) Determining performance levels; and (3) Developing full descriptions for the performance levels of the UIS Proficiency Scales. The performance level in step 2 is determined by deciding the number and writing the policy definitions for the performance levels of the UIS Proficiency Scales (UIS, 2017e). According to UIS (2017e), after performance levels of UIS proficiency scales for each grade and domain are defined, the next step is to link the scales with various national assessments (NAs) and cross-national assessments (CNAs) for the purpose of SDG 4.1.1 reporting. Due to the lack of a common assessment for SDG 4.1.1 reporting, they argue that it is not possible to statistically link the UIS proficiency scales with NAs or CNAs using test- or item-based linking methods (i.e., equating, calibration, projection, or statistical moderation). Rather, the linking is suggested through a content-based performance level expectation called the 'social moderation' or 'policy linking' (Buckendahl & Foley, 2015; Reckase, 2000; UIS, 2017e).

The two steps involved in the linking method, as suggested by UIS (2017e) are: (1) Evaluating alignment of performance level descriptors (PLDs), and (2) Setting socially moderated performance standards for NAs and CNAs. An example of the linking method for UIS Proficiency Scale with national and cross-national assessments is provided below.

Figure 2: Linking UIS Proficiency Scale with National and Cross-National Assessments: An Example



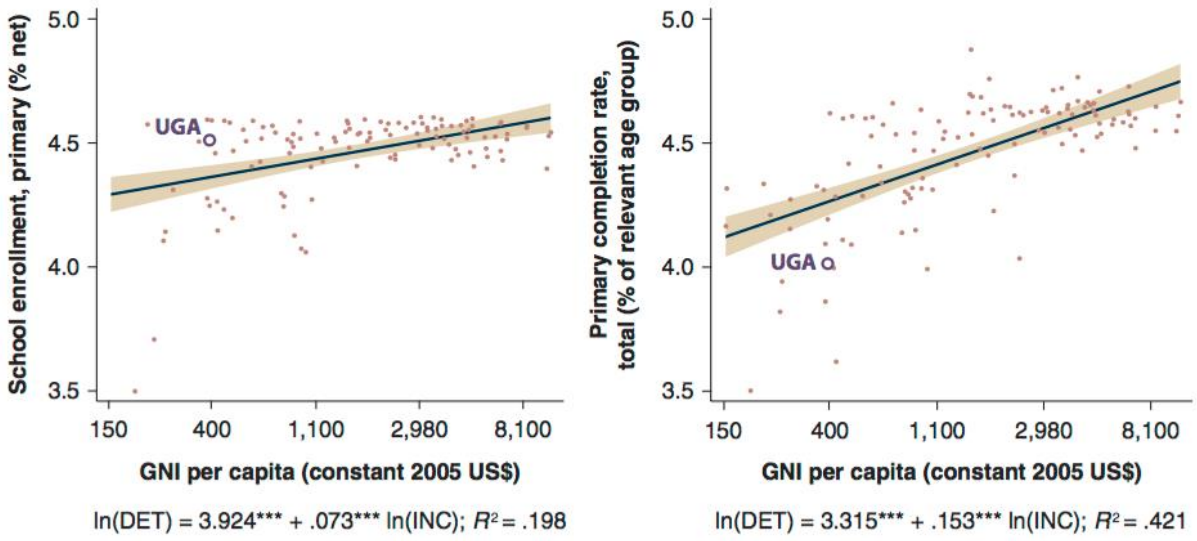
Source: UIS, 2017e

World Bank: Using GNI per capita for SDG Benchmarking

In 2013, the World Bank led the development of a framework to assess countries' abilities to achieve the SDGs, primarily using Gross National Income (GNI) per capita. Its experts reasoned that GNI per capita is highly correlated with SDG indicators due to its high correlation with influencing factors of SDGs, including per capita household incomes and tax revenues, both of which are used in spending in areas the SDGs impact. For each country, the framework consists of four steps: 1) benchmarking the current level of progress for each SDG relative to other countries, given GNI per capita, 2) projecting the country's SDG values by 2030 using a business-as-usual projection, 3) observing determinants of SDG outcomes and identifying ways of achieving outcomes higher than those discovered in the second step, including policy changes, and 4) discussing ways to expand spending on SDGs (Gable, Lofgren, & Osorio-Rodarte, 2015). This framework was first piloted in Uganda, and then applied, with available data, to nine other countries: Ethiopia, Jamaica, Kyrgyzstan, Liberia, Nigeria, Pakistan, Peru, Philippines, and Senegal. Among the questions that the framework sets to answer is what feasible targets for 2030 are possible within current income projections.

Cross-country, constant-elasticity regressions and their determinants on GNI per capita are used for benchmarking purposes, mostly in part for their simplicity and transparency to see how a country performs relative to others at its income level (Gable, Lofgren & Osorio-Rodarte, 2015). Compared to GDP per capita, GNI per capita is more closely related to a country's capacity to achieve the SDGs. Such a methodology allows one to observe whether a country is performing above or under predicted SDG values based on its capacity to achieve them. In its pilot of data from Uganda, shown in Figure 4, the fitted, straight line represents expected levels of school enrolment or completion for countries based on different GNI per capita.

Figure 3: Uganda – Primary School Net Enrolment and GNI per Capita (Left); Primary School Completion and GNI per Capita (Right)



Sources: WDI, EdStats.

Countries beyond the shaded areas as indicated in the graph show extreme over- or underperformance on these two targets relative to their GNI per capita. As can be seen for Uganda, net enrolment in primary school is significantly higher than expected; the opposite is true for completion rates at the primary level (Gable, Lofgren, & Osorio-Rodarte, 2015). With current performance established, and assuming the correlation between the SDGs and GNI per capita is strong, projections are then calculated to determine what countries will be able to achieve within a business-as-usual model. The projections suggest that Uganda's primary completion rate will increase, and the country in general will see significant yet moderate progress, though realizing the global goals remains beyond the 2030 mark. Though the two subsequent steps in this framework delve deeper into identifying areas where policy and spending can be explored and exploited for further change, the projections allow for the establishment of current benchmarks and those likely for the end of the SDG era in 2030, excluding the presence of accelerated growth.

Developing countries/Others

Ibrahim Index of African Governance

Currently in its 11th iteration, the Ibrahim Index of African Governance (IIAG) is an annual assessment of the quality of governance in each of the 54 African countries. Governance, according to IIAG's creating institution the Mo Ibrahim Foundation, is "the provision of the political, social and economic public goods and services that every citizen has the right to expect from his or her state, and that a state has the responsibility to deliver to its citizens" (Umar, 2017). The IIAG measures outputs and outcomes of policies, rather than declarations of intent, *de jure* statutes, and levels of expenditure.

Under this umbrella of governance, the IIAG framework includes education, within which uses eight indicators: Education Provision, Education Quality, Educational System Management, Human Resources in Primary School, Human Resources in Primary Schools, Primary School Completion, Secondary School Enrolment, Tertiary School Enrolment, and Literacy (Mo Ibrahim Foundation, 2017). The range of data for each indicator comes from both international and regional sources: for example, the indicator Education Provision assesses the extent to which the public is satisfied with how the government is addressing educational needs, taken from the Afrobarometer Survey developed in 37 African countries (Mo Ibrahim Foundation, 2017). The Education Quality indicator, on the other hand, derives its score from the Bertelsmann Transformation Index.

Drawing its data from 36 independent international sources, the IIAG holds them to two main criteria: it must be a proxy for governance which covers at least 33 of the 54 countries, and provides at least two years' worth of data for those countries since 2000 (Umar, 2017). Results are classified in three main types: score, rank and trend. Data is standardized in a 1-100 scale through min-max normalization for easy interpretation and comparison. Where data is missing, estimates derived from imputation are provided. Average Annual Trends (AAT) are also calculated to see either improvement or deterioration in each country's progress. With this structure, the IIAG can assess the region holistically, but also provide specific performance scores where needed. According to the IIAG, over the last decade (2007-2016), education has gained 3.6 points, but in the past five years has significantly slowed down (Mo Ibrahim Foundation, 2017).

Box 1: Inclusive education indicators relevant to Fijian context, ranked by frequency of survey response). From Sprunt et al. (2017).

1. Achievement (academic)
2. Participation in school and extracurricular activities
3. Independence (including responsibilities at home)
4. Employment
5. Enrolment
6. Participation in the wider community
7. Peer interaction and social skills
8. Self-esteem /confidence
9. Transition through levels of education
10. Family supportive of their child's education
11. Child's happiness and quality of life
12. Stakeholder involvement and approval
13. School attendance
14. Discrimination

Fiji and Benchmarking Disability-Inclusive Education

Providing inclusive, quality and free education for learners with disabilities has received more momentum from previous global efforts to improve education systems. With SDG 4 embracing this group for its 2030 Agenda, countries are building their approaches to monitoring the education of children with disabilities. For developing countries in particular, with relatively nascent organization, training, and resources in special education, developing standards with which to measure progress towards SDG 4 in relation to inclusive education must be thoughtfully considered.

In Fiji, a recent study sought to distinguish country stakeholders' priorities for indicators to measure disability-inclusive education, with fulfilling SDG 4 targets in mind. Surveying teachers, education officials, and parents, among others, the study found 14 indicators frequently repeated by Fijians as important to monitor success towards inclusive education (see Box 1). Of the 14 indicators, only 4 that were SDG-related would be measured: enrolment, achievement, transition through levels of education, and accessible school environments (Sprunt et al, 2017). Some of the indicators excluded from the overlap between SDG 4 and Fijian-recommended are viewed as reflective of Fijian societal

context, such as the participation in the wider community, and, in the study's view, must be prioritized.

In terms of measuring the SDG 4 inclusive education indicators, Fiji's Education Management System and/or household surveys are more than capable to collect data. However, what is relevant to the Fijian society is more difficult to measure at a national level. Specific measures have yet to be developed involving these indicators, though Fiji's Ministry of Education has committed to improving inclusivity. Possibilities to capture this data include periodic surveys, qualitative studies, and reviews of students' Individual Education Plans (Sprunt et al., 2017).

UNICEF: Data Must Speak Initiative (DMS)

At the global level, the Data Must Speak Initiative (DMS), led by UNICEF with funding from GPE, focuses on generating knowledge on strategies that work within a country's context to enhance school-level community participation and data usage for improving equity and learning (Bonnet & Kelly, 2017). The DMS builds upon data that governments collect through Education Management Information Systems (EMIS), but takes the information and creates user-friendly versions in the form of report cards and dashboards that are then distributed to all stakeholders involved in education, from ministry officials to parents (UNICEF, 2017). Implemented thus far in Nepal, Madagascar, the Philippines, Togo, and Zambia, the DMS indirectly contributes to the monitoring of SDG 4 through its focus on building countries' ability to use data.

Results from this initiative have led to various achievements in monitoring learning among participating countries. Notably, new indices have been developed in Nepal and the Philippines to help measure important components of SDG 4. In Nepal, a new equity index will allow the government to identify districts that are in need of additional support and the areas in which equity remains a challenge. In the Philippines, DMS facilitated the development of a teacher hardship index to reduce inequalities among teachers in schools around the country. Combining several "factors of hardship", including distance traveled to school and back, level of poverty, and access to electricity, the index is used to calculate the amount of "hardship allowance" a teacher is entitled to (Bonnet & Kelly, 2017).

Index of Development of Basic Education in Brazil

In March 2007, the government of Brazil launched its Education Development Plan to improve the quality of education in the country. In addition to emphasizing basic education – a change from its previous focus on higher education – was the historic creation of a synthetic indicator of education quality known as the Basic Education Development Index (IDEB) based on the average pass rate and average proficiency of each municipality in the country for its two national exams Prova Brasil and Saeb for students in 4^{th} and 8^{th} grade (Neri & Buchmann, 2008). Stemming from a national desire to set tangible objectives, the indicator uses the aforementioned factors as proxies for education quality in each level of education. Because the index is the product of both test scores and pass rates, it discourages schools from automatic promotion of children who are not learning as well as holding children back to boost learning scores (Bruns et. al, 2012).

To establish the levels of IDEB, the Saeb scales were harmonized with Brazil's 2003 PISA performance levels, which consisted on identifying which scores on the SAEB scale corresponded to a given performance in PISA and vice versa, influenced by the idea to make PISA a reference to

establish the goals of IDEB. Ranging on a scale from 0 to 10, the index value in Brazil was 3.8 in 2005, and its target for 6.0 was set for 2021 (Neri & Buchmann, 2008).

The data is collected from the IDEB are used as primary source for monitoring its education system, establishing public policies, educational research in different subareas and for the society, as well as for possible identification of municipalities for its cash-transfer programs (Carson et al., 2015). Particularly relevant is its ability to detect schools that are performing poorly and their changes in a longitudinal way. One instance of IDEB's impact can be seen in its ability to facilitate the implementation of teacher bonus programs across municipalities (Bruns et al., 2012). Thanks to IDEB, every single level of the Brazilian education system can benchmark how well its students are learning and how efficiently its school or school system is performing and set targets for their improvement (Bruns et al., 2012).

Other SDGs

The SDG Index and Dashboards

To assist countries in getting started with implementing the new SDGs, the SDG Index, developed by the Bertelsmann Stiftung and the Sustainable Development Solutions Network (SDSN) in 2016, creates for countries a measure of where they stand in relation to achieving the SDGs. Though not officially endorsed by the United Nations, the index aims to provide countries an easy-to-read snapshot of a country's starting point for measuring SDG progress, against which they can compare themselves to other countries and identify priority areas for action. The first iteration of the index covered only the 34 OECD countries; the third has expanded to include 157 of the 193 UN member countries (Sachs et al., 2017).

The SDG Index ranks countries on their initial status on each of the 17 SDGs, using available data on indicators for which countries have published data. Indicators were chosen based on several characteristics: global relevance and applicability to a broad range of country settings; statistical adequacy; timeliness; data quality; and coverage. While utilizing indicators suggested by the IAEG-SDGs that fit those criteria, the Bertelsmann Stiftung and the SDSN have also drew upon other resources, such as World Bank's World Development Indicator database and the Human Development Report by the UNDP, to include other indicators to monitor the SDGs. For example, the SDG Index includes scores on the Sustainable Nitrogen Management Index proposed by Zhang & Davidson (2016) as an indicator for SDG 2 (Zero Hunger). For Goal 4, the Index includes two indicators similar to the IAEG-SDGs provisional Tier 1 Indicators (IAEG-SDGs 2016): net primary school enrolment rate and the literacy rate of 15-24 year olds, both sexes (Sachs et al., 2016). They also use a country's PISA score as an indicator.

In order to compute the SDG Index, data from each indicator is ordered from worst to best. Worst values are created by first removing influential outliers and then identifying the next-worst value. This value is then applied to the bottom 2.5 percentile of the distribution. For creating the best score, the Index fulfills the phrase "leaving no one behind" by establishing absolutes to certain indicators, such as zero undernourishment and 100% school completion. Where no such absolute exists, the average of the top 5 values in a given indicator from sampled countries is assigned the best value. From there, countries are ranked from worst (0) to best (100) based on the aggregation of indicators using the arithmetic average. Turkey, for example, is ranked 67/157 on the Index with a

score of 68.5, meaning it is 68.5% of the way to the best possible outcome across the 17 SDGs (Sachs et al., 2017).

The Bertelsmann Stiftung and the SDSN believe that absolute thresholds are more appropriate due to the fact that most SDGs require absolute benchmarks to be achieved, thus each country featured in the SDG Index has its own country page which includes SDG Dashboards that represent the available data on SDG achievement across the 17 goals using a color-coded schema. Each goal is highlighted in green, yellow, or red, with red being the areas of greatest challenge. Green signifies that for this indicator the country is on a good path towards reaching an SDG and its targets or has (in some cases) already achieved the threshold consistent with SDG achievement. The SDG Index has found that no country is free from challenges in meeting the SDGs, though those challenges differ between countries and regions.

The SDG Index is not without its limitations, of which the 2016 report notes 4 initial ones, including the inclusion of non-official indicators and the exclusion of time series data. The 2017 report acknowledges them as driver for refinement; in this most recent report, the Index has revised not only its indicators, but also its methodology. It now covers 99 indicators across 157 countries, as well as adding indicators to address spillover effects (Sachs et al., 2017). Among its additions are HIV injections per 1,000 (SDG 3), E-waste by kg/capita (SDG 12), and mean protected area of freshwater sites (SDG 15). Because the data between the 2016 and 2017 reports have changed in terms of scope and types of indicators, they are not comparable enough to estimate trends to see how quickly countries are progressing on the SDGs.

Table 7: Summary of Existing Benchmark/Standardization Initiatives

Initiatives	Advantages	Disadvantages	Notes
The International Standard Classification of Education (ISCED)	• Serves as an instrument to compile and present education statistics both nationally and internationally. • Applied in statistics worldwide with the purpose of assembling, compiling and analyzing cross-nationally comparable data. • Is the reference classification for organizing education programs and related qualifications by education levels and fields. • Information compiled according to ISCED can be used for assembling statistics on many different aspects of education of interest to policymakers and other users of international education statistics.	• Needs updating as needed to better capture new developments in education systems worldwide. • Cannot be used to directly assess the competencies of individuals because there is no direct relationship between education programs or qualifications and actual educational achievement. • When classifying national education programs by ISCED levels, transition points between national programs and exit points into the labor market may not always coincide with transition points between ISCED levels.	• The ISCED is the standard framework used to categorize and report cross-nationally comparable education statistics that belongs to the United Nations International Family of Economic and Social Classifications.
UIS/OECD/EUROSTATA (UOE) Data	• Provides internationally comparable data (mostly at national level, with	• Is specific to OECD/EU region and is mostly not suitable for	• The preparation of the data collection

Initiatives	Advantages	Disadvantages	Notes
Collection	some insights at the subnational level) on key aspects of formal education systems, specifically on the participation and completion of education programs, as well as the cost and type of resources dedicated to education. • Participating countries co-operate to gather the information, to develop and apply common definitions and criteria for the quality control and verification of the data.	other countries that lack similar resources and capabilities.	tables is guided by the search for a common denominator between UNESCO-UIS, OECD and EUROSTAT.
Education and Training (ET) 2020	• Takes into consideration the whole spectrum of education and training systems with a lifelong learning perspective. • Covers all levels and contexts of learning. • The set benchmarks have comparable data and account for differing Member State situations. • Offers flexibility in measurement for Member States.	• Only seven benchmarks are established, and of those, only six are operational. • No metrics exist for harder to measure concepts (e.g., creativity). • Some indicators are not measured at all in participating Member States. • Poor design of Open Method of Communication (OMC) for benchmarking procedures.	• Many of the indicators were already in use to measure long-term EU policies. • No mandate exists for EU Member States to adopt ET 2020 benchmarks in their national standards. • Discussions surrounding benchmarking emphasize having a limited number to increase impact.
OECD' Indicators of Education Systems (INES)	• Addresses the issue of measuring the current state of education internationally. • Education indicators are organized thematically and each is accompanied by relevant background information. • Distinguishes actors in education systems: individual learners, instructional settings and learning environments, educational service providers, and the education system as a whole.	• The situation in any given country or economy may differ greatly from the average. • Sometimes the very low performers on PISA may include students who perform well relative to other student in their country/economy (OECD, 2016b). • It is also possible that a low-performing student in PISA may also be considered a high-performing student on a different assessment. • Fuels competition between OECD countries and emerging ones.	• Benchmarking in the OECD involves planning and defining the area of study, collecting, structuring and evaluating data, and reviewing and revaluating policy domains to identify effective approaches. • OECD relies on good arguments and a common value system to influence national

Initiatives	Advantages	Disadvantages	Notes
	• Groups indicators according to whether they are measures of learning outcomes for individuals and countries, policy levers or circumstances that shape these outcomes, or antecedents or constraints that set policy choices into context. • Identifies policy issues to which the indicators relate, with three major categories distinguishing between the quality of educational outcomes and educational provision, issues of equity in educational outcomes and educational opportunities, and the adequacy and effectiveness of resource management. • Follows a rigorous structure plan to set benchmarks. • Little competition between Member States, leading to smooth benchmarking process.		policymakers.
Regional Assessments	• Involve similar methodologies to measure students' learning.	• Different levels of minimum proficiency, target population, and incomparable iterations of exams lead to reliability issues in their comparison.	• Efforts are being made to monitor SDG 4 with regional assessments. • Efforts to establish comparability among regional assessment are under way.
Qualifications Framework: National Qualifications Framework (NQF) and Regional Qualifications Framework (RQF)	• Introduces different levels of standards which describe the characteristics and context of learning as is expected at each level. • RQFs as tools for supporting cross-border mobility of learners and workers and acting as a means for fair and transparent recognition of qualifications. • Aims to improve not only the quality of worker qualifications but also their relevancy in the modern workplace.	• Global comparisons of QFs are not yet operational; while regional ones are becoming more thorough. • Initial expectations are too high in terms of both what can be achieved and how quickly the benefits of introducing a QF are likely to become apparent.	• NQFs are now being implemented or developed in over 150 countries as of 2015

Initiatives	Advantages	Disadvantages	Notes
	• Within NQFs are set standards against which to benchmark the development of TVET skills, contributing to monitoring an important new aspect of SDG 4. • QFs can support in measuring the impact of SDG 4 by addressing the list of indicators established by SDG 4. • RQFs can set common standards for competences. • New learning outcomes and setting future targets in a systematic way have been inspired by NQFs in many European countries. • Provides education authorities the ability to develop more relevant curricula for learners to develop the skills to succeed. • Having a standard set of qualifications that respond to and benchmark increasingly globalized industries, can catalyze a more systematic approach to skills development in education.		
Global Alliance to Monitor Learning (GAML)	• Aims to bring together national education authorities, assessment agencies, citizen-led initiatives and the international education community, including donors, to ensure that countries have the high-quality data needed to improve the learning outcomes of all and to track progress globally. • Intends to target its resources and expertise to assist countries in achieving SDG 4. •At the national level, works with partners to develop tools, standards and guidelines to help countries who do not have national learning assessments to develop one, and for those that do, to improve efficiency and efficacy in utilizing that data. • At the global level, seeks to establish a common framework and data	• At an early phase of its development. • Establishing global comparisons entails significant methodological and statistical efforts.	• Conceptualized by UIS in early 2017 with two basic objectives: 1) to support national strategies for learning assessment, and 2) to ensure international reporting on the SDGs by all UN member states.

Initiatives	Advantages	Disadvantages	Notes
	validation process for quality global reporting.		
World Bank: Using GNI per capita for SDG Benchmarking	• Compared to GDP per capita, GNI per capita is more closely related to a country's capacity to achieve the SDGs. • Allows one to observe whether a country is performing above or under predicted SDG values based on its capacity to achieve them. • The subsequent steps in this framework delve deeper into identifying areas where policy and spending can be explored and exploited for further change. • The projections allow for the establishment of current benchmarks and those likely for the end of the SDG era in 2030, excluding the presence of accelerated growth.	• There is much more about countries performance than basing solely on GNI per capita. • Makes comparisons based on economic principles, which may not be appropriate for measuring or benchmarking certain SDG 4 indicators (e.g. citizenship).	• Cross-country, constant-elasticity regressions and their determinants on GNI per capita are used for benchmarking purposes, mostly in part for their simplicity and transparency to see how a country performs relative to others at its income level.
Fiji Benchmarks in Inclusive Education	• Some SDG indicators cover the needs of monitoring inclusive education. • Fiji can measure SDG 4 indicators using information management systems and household surveys.	• Developing nations may not yet have the capacity to monitor children with disabilities effectively. • What is important to Fijian society may be more difficult to measure at the national level.	• A 2017 survey of Fijian education stakeholders found 14 indicators seen as important to monitor inclusive education, 4 of which overlap with current SDG 4 indicators.
UNICEF: Data Must Speak Initiative (DMS)	• Focuses on generating knowledge on strategies that work within a country's context to enhance school-level community participation and data usage for improving equity and learning. • Builds upon data that governments collect through Education Management Information Systems (EMIS). • Indirectly contributes to the monitoring of SDG 4 through its focus on building countries' ability to use data. • Results from this initiative have led	• All countries might not have good EMIS systems.	• New indices for measuring equity in Nepal and Philippines have been developed based on data from DMS.

Initiatives	Advantages	Disadvantages	Notes
	to various achievements in monitoring learning among participating countries.		
Ibrahim Index of African Governance	• Measures outputs and outcomes of policies, rather than declarations of intent, de jure statutes, and levels of expenditure. • Results are classified in three main types: score, rank and trend. • Can assess the region holistically, but also provides specific performance scores where needed. • Indicators are chosen based on applicability and availability of data for the best relevance to the region.	• Score, rank and trend might not give the complete picture as it ignores countries' starting points. • Its use of international measurement tools incurs their biases in data collection and analysis. • Estimations and Average Annual Trends utilized for missing data and projections are subject to bias.	• Currently in its 11th iteration, the Ibrahim Index of African Governance (IIAG) is an annual assessment of the quality of governance in each of the 54 African countries.
Index of Development of Basic Education (IDEB) in Brazil	• Covers all municipalities of Brazil, providing information on every level of the education system to establish targets • Discourages promotion or holding back of students based on its design. • Can assess the region holistically, but also provides specific performance scores where needed. • Identified low-performing areas are targeted with financial and supportive resources.	• May not be widely distributed to local communities for accountability purposes. • Momentary fluctuations in a single school may affect the IDEB score for the whole municipality. • No method to account for the possibility that teachers are coaching students on how to take the Prova Brasil	• Ranging on a scale from 0 to 10, the index value in Brazil was 3.8 in 2005, and its target for 6.0 was set for 2021. • The data is collected from the IDEB are used as primary source for monitoring its education system, establishing public policies, educational research in different subareas and for the society, as well as for possible identification of municipalities for its cash-transfer programs.
SDG Index and Dashboards	• Aims to provide countries an easy-to-read snapshot of a country's starting point for measuring SDG progress, against which they can compare themselves to other countries and identify priority areas for action. • Ranks countries on their initial	• Not officially endorsed by the United Nations. • The inclusion of non-official indicators and the exclusion of time series data. • Because the data between the 2016 and 2017 reports have changed in terms of scope and	• Have also drew upon other resources, such as World Bank's World Development Indicator database and the Human Development Report by the UNDP, to

Initiatives	Advantages	Disadvantages	Notes
	status on each of the 17 SDGs, using available data on indicators for which countries have published data. • Indicators were chosen based on several characteristics: global relevance and applicability to a broad range of country settings; statistical adequacy; timeliness; data quality; and coverage. • Also uses a country's PISA score as an indicator.	types of indicators, they are not comparable enough to estimate trends to see how quickly countries are progressing on the SDGs.	include other indicators to monitor the SDGs.

III. A Conceptual Framework for Benchmarking

Recent benchmarking efforts at both the GAML and UIS have centered on establishing a proficiency scale linking with national and cross-national assessments through the process of social moderation/policy linking to measure progress on SDG 4.1.1. Despite the statistical challenge this effort presents, taking advantage of current measurement tools that are already increasingly used presents an opportunity to advance their usefulness in measuring learning. It is an important step in the right direction towards advancing the world's knowledge on benchmarking at a global level.

However, there are important considerations to make. In Treviño & Ordenes' (2017) four time-bound strategies for assessing SDG 4, the three mid- to long-term strategies (including the development of a worldwide assessment) all greatly reduce the external validity in representing national curricula. Not aligning metrics to national policy and curricula will reduce their use and usefulness in informing policy development and supporting classroom interventions as they diverge from countries' needs and priorities. Such effects could go in at least two ways: either the measurement of SDG 4 fails to accurately capture and influence learning, or that countries will increasingly push towards a homogenized curriculum (and with that the erasure of culture in curriculum). While some suggest that ignoring the impact of culture on learning and moving beyond static measures allows for the definition/creation of a global curriculum (and therefore global benchmarks), culture stills plays a significant part in the education system to be ignored. Establishing global content standards, even with experts meeting to determine them, elicits further questions of which experts are chosen and what/who they represent, as well as what might be ignored or forgotten in benchmark development.

Another consideration is understanding how to account for countries' different starting points. As mentioned in previous sections, certain indices are biased towards developed countries, and equating economic prosperity to high educational outcomes is not always a strong relationship. Ensuring that countries like Namibia won't be both disadvantaged or misrepresented when compared to a country like France in progress or final reports on SDG attainment is fair by taking into account both status and rates of change.

Status is usually defined as the level of attainment or coverage for a particular indicator. However, using status to compare countries does not reflect a country's performance, as a country can be increasing, decreasing, or stagnating in its attainment or coverage level. Luh et al. (2016) term these

as progression, regression, and stagnation, respectively. The second approach would be to assess trends in status over time, i.e., rates of change. Rates of change provides information on whether a country is moving towards a target and how fast of slow is the improvement over time. It is essential to note that status cannot be completely disregarded as it affects a country's rate of change. As Luh et al. (2016) explain, this is because "countries at very high levels of coverage can only make small improvements as they approach 100% because the remaining unserved become more difficult to reach, and countries at very low levels of coverage make little initial progress as the systems, policies, and infrastructure required are not yet in place."

While both status and rates of change are important measures of improvement or regression on targets, individually they fail to sufficiently consider the differences between countries' starting points. Indeed, this method of measurement by either status or rate of change alone has been viewed by some scholars as an inefficient and unfair method in comparing countries that have vastly differing histories, capacities, and economic situations (Fukudu-Parr & Randolph, 2014). This draws upon the previous view of differentiated targets, but with a specific methodology to achieve that. Therefore, instead of blindly benchmarking all countries to one another, we suggest using the principle of frontier analysis to fairly compare countries on their respective progress on the SDG 4. Below is a brief description of frontier analysis.

Frontier analysis is a non-parametric method that applies the principles of data envelopment analysis (DEA) to evaluate and compare the efficiency of decision making units (DMUs) such as hospitals, schools, and banks and to define a "best-practice frontier" or "benchmark" for operations management (Cook et al, 2014; Luh et al., 2016). The efficiency of a DMU is then calculated as the ratio of its distance from the best-practice frontier.

To assess progress, linear rates of change are calculated to describe how the metric of interest changes with time. The criteria for data inclusion depends on the overall objective of the study as well as data availability. Once rates of change are calculated, these rates need to be associated to a corresponding metric of interest. The performance index is calculated by using the following formula:

Index = (country rate – minimum rate)/(maximum rate – minimum rate).

The maximum rate is defined as the maximum possible rate achievable, or benchmark rate, at the coverage level of the country analyzed and is calculated from the best-practice frontier, and the minimum rate is defined as zero (no progress) (Luh et al., 2016).

A set of 43 indicators, including the 11 global indicators recommended by the Inter-Agency and Expert Group on SDG indicators have been approved by the Technical Cooperation Group for SDG4-Education 2030 Indicators. However, establishing benchmarks for each of these 43 indicators is not an optimal option, particularly for many developing countries that lack the resources in monitoring their educational goals. Therefore, we suggest using selected indicators for this benchmarking process, which should be determined by existing team of experts working in this area.

An example for one indicator is provided below. Obviously, it is important to use proper standardized definitions (such as of ISCED) while measuring status and rates of changes of the indicator being considered. In this framework, three points of time are suggested: 2020, 2025 and 2030, and three measures are considered: average status, trend in status, and the performance

index. The performance index of a particular country (simply called 'index' in the table below) is measured using the formula stated above. Two schemes are provided below. The first scheme shows a country's performance for various indicators at different points of time. The second scheme shows information on how countries are performing on a particular indicator at various points of time. Either or both of these schemes could be considered depending on the relevance, need and data availability.

Scheme 1: Country A's Performance: By Indictors

Indicators/Time	2020			2025			2030		
	Avg. Status	Trend in Status	Index	Avg. Status	Trend in Status	Index	Avg. Status	Trend in Status	Index
Proportion of children and young people in grades 2/3 achieving at least a minimum proficiency level in reading for female									
Proportion of children and young people at the end of primary achieving at least a minimum proficiency level In reading for female									
Proportion of children and young people at the end of lower secondary achieving at least a minimum proficiency level in reading for female									

Scheme 2: Indicator 1 (By countries): Proportion of children and young people in grades 2/3 achieving at least a minimum proficiency level in reading for female

Indicators/Time	2020			2025			2030		
	Avg. Status	Trend in Status	Index	Avg. Status	Trend in Status	Index	Avg. Status	Trend in Status	Index
Country A									
Country B									
Country C									

Indicator 1(Example): Proportion of children and young people in grades 2/3 achieving at least a minimum proficiency level in reading for female

Summary and Discussion

An OECD report states, "Making SDG 4 a reality will transform lives around the globe. Education is so central to the achievement of a sustainable, prosperous and equitable planet that failure to achieve this particular SDG puts at risk the achievement of the 17 SDGs as a whole" (OECD, 2017a). Moreover, the recent World Development Report make clear that education is also a foundation block for nearly every other SDG: it saves lives, improves health, and fosters shared understanding and values (World Bank, 2018). Achieving SDG4 will therefore be instrumental in realizing the broader aspirations of the SDG agenda.

As a consequence, the international community will need to invest substantially in achieving this necessary condition in the global fight against poverty and the achievement of a sustainable planet for all (OECD, 2017a). For education to deliver its full potential, participation rates have to dramatically improve, learning needs to become a lifelong pursuit and education systems need to fully embrace sustainable development (UNESCO, 2016a). Therefore, monitoring SDG 4 is extremely crucial in the global community to move towards this common goal. And, in this monitoring process, benchmarking is one of the options being considered by the experts and agencies working in this area.

However, as the review highlights, benchmarking is a complex process. Table 8 summarizes common benchmarking pitfalls and the conditions which lead to them. Perhaps the biggest conundrum facing global benchmarking is the technical difficulty of its implementation. One simple reason is that countries do not have the same standards for levels of education, what they learn at a particular age, and what goes into the curriculum. The flexibility in the different levels within which countries can measure their progress in SDG 4 and the other development goals, while avoiding a homogenization of schooling, may result in too much variability in global benchmarking. Even existing efforts to measure learning through cross-national assessments like those from LLECE and SACMEQ, despite employing similar methodologies in their assessment design fall victim to the same issue of incomparability because of context. Such a limitation in context may make international comparison extremely challenging.

Table 8: Pitfalls in international benchmarking

Cluster	Background conditions	Resulting in pitfall
Choice of benchmarking approach	International benchmarking can only be done on a consensual basis, no coercion,	(1) Mismatch: Choice for hierarchical, disciplinary standards and/or results (functional) benchmarking without corresponding coercion mechanisms
Selection of criteria, indicators	Multitude of relevant criteria and objectives (inherent to complex policies and policy systems)	(2) Pick-and-mix approach to benchmarking
	Disagreement on criteria due to national diversity in preferences	(3) Construction of common objectives is disguised as benchmarking
	Choice of peers/partners is institutionally determined	(4) Inclusion of irrelevant benchmarking partners
	Data availability problems	(5a) Over-reliance on indicators that are easily available, but may not be relevant to the criteria at hand (5b)Over-reliance on quantitative data

Cluster	Background conditions	Resulting in pitfall
Policy transfer	Complexity of policies and policy systems, limited amount of indicators taken into account	(6a) Uninformed transfers
	Complexity of policy systems, and diversity in national institutional contexts	(6b) Incomplete transfers
	Diversity of preferences	(6c) Inappropriate transfers

Source: Groenendijk, 2009

Pointers/Recommendations

Despite these challenges, it is worthwhile to note that the SDGs do aim to be more inclusive in what qualifies as sustainable development and promoting country ownership of SDG progress. The global desire to measure learning, with the galvanizing belief that better data will lead to more students achieving, is found at the heart of many of the current initiatives studied in this review. In this context, we conclude this paper by presenting some initial pointers and recommendations relating to benchmarking efforts of SDG4 for consideration in moving forward.

Global Vs. Other levels

Setting benchmarks for all levels of SDG 4 and the monitoring of those will definitely overwhelm UIS as well as the participating countries. Discussions are still ongoing as to which level should be focused on. The 2016 GEM report recommends that rather than overhauling the ways in which data is collected in education, better coordination between agencies and more resources to implement plans would be more effective in the changes needed to monitor the 2030 education agenda (UNESCO, 2016a). Some scholars argue that the agenda itself is not actually universal because of how targets are set within the goals that direct attention to developing nations more than developed ones, such as nutrition issues in Goal 2 being dominated by malnutrition and not shared with the equally threatening problem of obesity (van Bergeijk & van de Hoeven, 2017).

Countries in a given region tend to have common education contexts, thus setting benchmarks at the regional level may have better applicability and political consensus among them rather than global goals. For regional reporting, existing mechanisms, such as the Regional Economic Commissions, should work as a foundation to foster dialogue and knowledge-sharing among similar regions. Regional monitoring processes can also negotiate what is being measured at the national and global levels, especially if organizations are already subsidiaries of international organizations. Thematic reporting could be left to the coordination among specialized organizations, universities, and even businesses, which may have access to data.

One of the recommendations by the OWG that seems meaningful is each government setting its own national targets inspired by those at the global level due to country context (King, 2015). It is important that countries have ownership of the SDGs to realize the necessary changes. Nations may choose a combination of the Global Reporting Indicators and the Complementary National Indicators to harmonize global and national reporting. As can be seen through benchmarking efforts at the regional (EU and OECD) and national level (Brazil, Fiji) in this paper, every region and nation, no matter the number of similarities, have their own interests when it comes to their specific development. It is unlikely that all of those interests will align at a large scale. Vandemoortele (as

cited in van Bergeijk & van der Hoeven, 2017) suggests that global assessment needs to pay more attention to how the global targets make a difference at the national and sub-national levels.

It is important that countries have accessible, comprehensive and communicable data and can enhance the monitoring of progress within the SDGs at the local and subnational levels of government. Dialogues between ministries and statistic agencies must improve to achieve that. Because national governments are expected to integrate the global SDG 4 commitments into national education development efforts, the establishment of appropriate intermediate national/ local benchmarks seems like a good option, where the intermediate benchmarks for each target can serve as quantitative goalposts for review of overall progress vis-à-vis the longer-term goals (UNESCO, 2017). Combining those with regional benchmarks seems to be an effective manner with which to monitor progress towards SDG 4.

Absolute Vs. Relative

While some initiatives studied in this review suggest that the SDGs require absolute benchmarks, others suggest that this is an ineffective practice for monitoring a country's progress alone. For example, the OECD examines the distance to travel in order to reach each target level that involves determining levels of achievement on each target level. The level was pre-determined in the 2030 Agenda, either as a fixed value or as a relative improvement on a country's starting position (OECD, 2017c). Likewise, even before the SDGs, global goals and targets were expressed in either absolute terms or as combined relative and absolute benchmarks. Some scholars argue that neither type of benchmark taken alone provides the full picture of a country's progress or situation (Vandermoortele & Dalemonica, 2010). A combination of relative and absolute benchmarks arguably constitutes the best guarantee against possible biases in setting global targets. Therefore, our recommendation is not to get confined with one method. It will depend on a multiple factors, such as the how the target is set, whether it is clearly quantifiable, to what extent the initial position is important, and so on. Therefore, an indicator will have to be expressed using either or both terms, but should also have other methods considered as well. Experts will need to decide on this depending on the indicators and what is currently available to measure them.

Caution on measures/methods that consider only economic aspect

Another trend identified in this review is the tendency for progress to be measured in terms of economic progress. As shown by Dill and Gebhart (2016), many of the current indices used to track SDG progress inherently favor developed countries over developing countries. While certainly applicable in development as a whole, economic growth is not the only barometer against which countries and its individual citizens change and "develop" in education. Especially considering the rise of BRICS and MINT countries and the ways in which they are developing, benchmarking countries based on old ideas of development must give way to transformational change that is structural, institutional and normative. The push for more qualitative data by countries and institutions alike is also a promising start, and strategies to incorporate them into SDG 4 reporting must consider how they can better measure components of education, such as quality.

Differences in starting points and national capabilities

Another important aspect to consider is to use measures that go beyond the assessment of status and rates of changes alone. One such approach is frontier analysis that identifies benchmark rates

using the rate of the historically best performing country among those at a similar level of coverage or attainment. Setting one-size-fits-all quantitative and time-bound targets without taking account of differences in starting points and national capacities might not be realistic (Clemens, Kenny, & Moss, 2007) and could be unfair to countries that start farther from the target and face larger resource and other capacity constraints (Easterly, 2009; Fukuda-Parr, Greenstein, & Stewart, 2013). The benchmarking option should therefore consider different starting points or levels of development and available resources to avoid such risks. It may not be possible to capture all the data required of the SDGs with this process, but it can be a helpful start. The challenge with monitoring learning is huge, but as SDG 4 continues to influence and is influenced by the education landscape, countries and partners must work together on methodological development, sharing lessons learned and implementing new global measures in order to advance the measurement agenda.

References

Adams, B., & Judd, K. (2016). 2030 Agenda and the SDGs: Indicator framework, monitoring and reporting. Global Policy Watch Briefing #10. New York: Global Policy Forum.

Association for the Development of Education in Africa (ADEA) (2017). Zambia approves the Learning Assessments Systems Evaluation Framework developed by ADEA and NALA. [Press release]. Retrieved from: http://www.adeanet.org/en/news/zambia-approves-the-learning-assessments-systems-evaluation-framework-developed-by-adea-and.

Australian Council for Educational Research (ACER) (2014). The Latin-American Laboratory for Assessment of the Quality of Education: Measuring and comparing educational quality in Latin America. Assessment GEMS Series No. 3. ACER Centre for Global Education Monitoring. Available at https://www.acer.org/files/AssessGEMs_LLECE.pdf.

Australian Council for Educational Research (ACER) (2017). "Setting benchmarks on the UIS Reporting Scales", discussion paper for the 4th meeting of GAML. Montreal: UNESCO Institute for Statistics (UIS).

Birdsall, N., Bruns, B., & Madan, J. (2016). "Learning Data for Better Policy: A Global Agenda." CGD Policy Paper. Washington, DC: Center for Global Development. http://www.cgdev.org/sites/default/ les/learning-data-better-policy.pdf.

Bishop, M. (2016). Rethinking the Political Economy of Development Beyond 'The Rise of the BRICS'. Sheffield Political Economy Research Institute Paper No. 30. Available at http://speri.dept.shef.ac.uk/wp-content/uploads/2016/07/Beyond-the-Rise-of-the-BRICS.pdf.

Bonnet, G., & Kelly, D. (2017). Supporting effective education systems: The Data Must Speak initiative. Education for All Blog. Retrieved from: http://www.globalpartnership.org/blog/supporting-effective-education-systems-data-must-speak-initiative.

Broomé, A. and J. Quirk (2015). 'The Politics of Numbers: The Normative Agendas of Global Benchmarking.' Review of International Studies 41(5): 813-818.

Bruns, B., Evans, D., & Luque, J. (2012. Achieving World-Class Education in

Brazil : The Next Agenda. Washington, DC: World Bank. Retrieved from https://openknowledge.worldbank.org/handle/10986/2383.

Carson, L., Noronha, J.V., & Trebilcock, M.J. (2015). Held Back: Explaining the Sluggish Pace of Improvement to Basic Education in Developing Democracies—The Cases of India and Brazil. *Journal of Poverty Alleviation and International Development, 6*(2), 2-46.

Buckendahl, C. W., & Foley, B. P. (2015). Policy linking as cut score moderation: Considerations for practice. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Chakroun, B., & Daelman, K. (2015). Developing world reference levels of learning outcomes: potential and challenges. In: UIL; UNESCO, ETF; Cedefop (eds). Global inventory of regional qualifications frameworks, Volume I: thematic chapters. Hamburg: UIL.

Chakroun, B., & Ananiadou, K. (2017). Learning Outcomes World Reference Levels: Global Solutions for Global Challenges. In: UIL; UNESCO, ETF; Cedefop (eds). Global inventory of regional qualifications frameworks, Volume I: thematic chapters. Hamburg: UIL.

Cizek, G.J. (2006). Standard Setting. In Downing, S. & Haladyna, T. (Eds.). *Handbook of Test Development* (pp225-258). Mahwah: Lawrence Erlbaum Associates.

Clemens, M., Kenny, C., & Moss, T. (2007). The Trouble with the MDGs: Confronting Expectations of Aid and Development Success. *World Development*, *35* (5), 735-751.

Cook WD, Tone K, Zhu J. Data envelopment analysis: prior to choosing a model. Omega. 2014; 44:1–4.

De la Mothe, E., Espey, J., & Schmidt-Traub, G. (2015). Measuring Progress on the SDGs: Multi-level Reporting. Global Sustainable Development Report 2015 Brief. Available at: https://sustainabledevelopment.un.org/content/documents/6464102-Measuring%20Progress%20on%20the%20SDGs%20%20%20Multi-level%20Reporting.pdf.

Dill, A., & Gebhart, N. (2016). Redundancy, Unilateralism and Bias beyond GDP – results of a Global Index Benchmark. Munich Personal RePEc Archive Paper No. 74268. Munich: Basel Institute of Commons and Economics. Retrieved from: https://mpra.ub.uni-muenchen.de/74268/.

Easterly, W. (2009). How the Millennium Development Goals are Unfair to Africa. *World Development*, *37* (1), 26-35.

European Commission (2017a). ET2020 benchmarks: country evaluations (draft). European Commission. IEG on Education and Training Evidence Monitoring – 2nd Meeting 24th October 2017.

European Commission (2017b). ET2020 benchmarks: summary of country evaluations and overall conclusions (draft). European Commission. 2nd Meeting 24th October 2017.

European Commission/EACEA/Eurydice (2016). Structural Indicators for Monitoring Education and Training Systems in Europe – 2016. Eurydice Background Report to the Education and Training Monitor 2016. Eurydice Report. Luxembourg: Publications Office of the European Union.

European Union (2017a). Education and Training Monitor 2017. Luxembourg: Publications Office of the European Union.

European Union (2017b). Sustainable development in the European Union: Monitoring Report on Progress Towards the SDGs in an EU Context 2017 Edition. Luxembourg: Publications Office of the European Union.

Fukuda---Parr, S., Greenstein, J., & Stewart, D. (2013). How should MDG Implementation be Measured: Faster Progress or Meeting Targets?. *World Development*, *41* (1), 19-30.

Fukuda-Parr, S. & Randolph, S. (2014). Achievement Possibilities Frontier – a methodology for setting nationally specific benchmarks to meet a universal target: A proposal for how to set universal targets in the Post-2015 Agenda, adapted to the different levels of development and resources of difference countries. Storrs: University of Connecticut. Retrieved from: https://serfindex.uconn.edu/2014/02/27/achievement-possibilities-frontier/.

Gable, S., Lofgren, H., Osorio Rodarte, I. (2015). Trajectories for Sustainable Development Goals: Framework and Country Applications. Washington, DC: International Bank for Reconstruction and Development and the World Bank.

Global Partnership for Education (GPE) (2017). GPE Results Report 2015/2016. Washington, DC: The Global Partnership for Education.

Groenendijk, N. (2009, April). EU and OECD Benchmarking and Peer Review Compared. Paper presented at the Third European Union Centre of Excellence Annual Conference, Halifax, Canada.

Hambleton, R. M. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In Hansche, L. N. (Ed.). *Handbook for the development of performance standards* (pp. 87-114). Washington, DC: Council of Chief State School Officers.

Hambleton, R. M., Pitoniak M.J. & Copella, J. M. (2012): Essential Steps in Setting Performance Standards on Educational Tests and Strategies for Assessing the Reliability of Results. In Cizek (Ed.) *Setting Performance Standards Foundations, Methods and Innovations*. (pp 93-125) New York: Routledge.

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., van Cappelle, F., Paviot, L., & Vellien, J. (2010). SACMEQ III Project Results: Pupil achievement levels in reading and mathematics. Working Document No. 1. Gaborone: SACMEQ.

IAEG-SDGs (2016). Provisional Proposed Tiers for Global SDG Indicators as of March 24, 2016. New York, Inter-agency and Expert Group on Sustainable Development Goal Indicators.

International Commission for Financing Global Education Opportunity (Education Commission). 2016. "The Learning Generation: Investing in education for a changing world." New York, NY.

Kamens, D. (2013). Globalization and the Emergence of an Audit Culture: PISA and the search for 'best practices' and magic bullets. In H.D. Meyer & A. Benavot edits, *PISA, Power, and Policy: the emergence of global educational governance* (pp. 117-140). Oxford: Symposium Books.

King, K. (2015). The Global Targeting of Education and Skill: Policy History and Comparative Perspectives. Working Paper #9. Geneva: NORRAG. Retrieved from: http://old.norrag.org.www438.your-server.de/fileadmin/Working_Papers/Working_Paper__9_King.pdf.

Learning Assessment Capacity Index (LACI) (2017). UIS. Retrieved from: http://uis.unesco.org/apps/visualisations/laci/map.html#

Luh, J., Cronk, R,, & Bartram, J. (2016). Assessing Progress towards Public Health, Human Rights, and International Development Goals Using Frontier Analysis. *PLoS ONE*11(1): e0147663. https://doi.org/10.1371/journal.pone.0147663

Martens, J. (2015) Benchmarks for a truly universal post-2015 agenda for sustainable development. *Regions & Cohesion, 5*(1), 73–93.

Mo Ibrahim Foundation (2017). 2017 Ibrahim Index of African Governance Index Report. Dakar: Mo Ibrahim Foundation. Retrieved from: http://s.mo.ibrahim.foundation/u/2017/11/20124505/2017-IIAG-Report.pdf.

Montoya, S. & Hastedt., D. (2017 September 28). "News from Hamburg: Big Steps Forward Towards Reliable Metrics to Harmonise Learning Assessment Data Globally by Silvia Montoya and Dirk Hastedt." Retrieved from: http://www.norrag.org/news-hamburg-big-steps-forward-towards-reliable-metrics-harmonise-learning-assessment-data-globally-silvia-montoya-dirk-hastedt/.

National Governors Association (NGA) (2008). Benchmarking for Success: Ensuring U.S. Students Receive a World-Class Education. Washington, DC: NGA, Council of Chief State School Officers (CCSSO), and Achieve.

Neri, M. C., and Buchmann, G. 2008. The Brazilian Education Quality Index (Ideb): measurement and incentive upgrades. In: LACEA/LAMES, 2008, Rio de Janeiro.

OECD, Eurostat and UNESCO Institute for Statistics (2015). ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications, OECD Publishing, Paris. Available at: http://dx.doi.org/10.1787/9789264228368-en.

OECD (2016a). Education at a Glance 2016: OECD Indicators, OECD Publishing, Paris.

OECD (2016b). Low-Performing Students: Why They Fall Behind and How to Help Them Succeed. OECD Publishing: Paris.

OECD (2017a). Education at a Glance 2017: OECD Indicators, OECD Publishing, Paris. Available at: http://dx.doi.org/10.1787/eag-2017-en

OECD (2017b). OECD Handbook for Internationally Comparative Education Statistics: Concepts, Standards, Definitions and Classifications, OECD Publishing, Paris. Available at: http://dx.doi.org/10.1787/9789264279889-en.

OECD (2017c). Measuring Distance to the SDG Targets: An assessment of where OECD countries stand. Paris: OECD Publishing.

Programme d'Analyse des Systèmes Educatifs (PASEC) (2015). PASEC2014 Education System Performance in Francophone Sub-Saharan Africa: Competencies and Learning Factors in Primary Education. Dakar: Programme d'Analyse des Systèmes Educatifs de la CONFEMEN.

Reckase, M. D. (2000). The evaluation of the NEAP achievement levels setting process: A summary of the research and development efforts conducted by ACT. Iowa City, IA: ACT, Inc.

Sachs, J., Schmidt-Traub, G., Kroll, C., Durand-Delacre, D., & Teksoz, K. (2016): SDG Index and Dashboards - Global Report. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN).

Sachs, J., Schmidt-Traub, G., Kroll,C., Durand-Delacre, D. & Teksoz, K. (2017): SDG Index and Dashboards Report 2017. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN).

Sandefur, J. (2016). Internationally Comparable Mathematics Test Scores for Fourteen African Countries. CGD Working Paper 444. Washington, DC: Center for Global Development. Retrieved from: http://www.cgdev.org/publication/math-scores-fourteen-african-countries.

Sparapani, E.F., Perez, D.C., Gould, J., Hillman, S., & Clark, L. (2014). A Global Curriculum? Understaning Teaching and Learning in the United States, Taiwan, India, and Mexico. *SAGE Open,* pp. 1-15.

Sprunt, B., Deppeler, J., Ravulo, K., Tinaivunivalu, S., & Sharma, U. (2017). Entering the SDG era: What do Fijians prioritise as indicators of disability-inclusive education? *Disability and the Global South, 4*(1),1065-1087.

United Nations (2014). Open Working Group Proposal for Sustainable Development Goals. New York, United Nations.

United Nations (2015). Transforming our world: The 2030 agenda for sustainable development. 2nd August 2015. New York: United Nations.

UNESCO (2016a). Education for people and plant: creating sustainable future for all. Global Education Monitoring Report. United Nations Educational, Scientific and Cultural Organization, Paris.

UNESCO (2016b). Education 2030 Incheon Declaration and Framework for Action. Paris/New York/Washington, DC, Geneva, Switzerland, UNESCO/UNDP/UNFPA/UNHCR/UNICEF/UN Women/World Bank/International Labour Organization.

UNESCO (2017). Unpacking Sustainable Development Goal 4 Education 2030. Guide. United Nations Educational, Scientific and Cultural Organization, Paris. Available at: United Nations Educational, Scientific and Cultural Organization, Paris.

UNESCO Institute for Statistics (UIS) (2012). International Standard Classification of Education: ISCED-2011, UNESCO Institute for Statistics, Montreal. Available at: http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

UNESCO Institute for Statistics, OECD and Eurostat (2016a). UOE Data Collection on Formal Education, Manual on Concepts, Definitions and Classifications, OECD and Eurostat, Montreal, Paris, Luxembourg. Available at: https://circabc.europa.eu/sd/a/849a866e-d820-4006-a6af-21cb1c48626b/UOE2016manual_12072016.pdf.

UNESCO Institute for Statistics, OECD and EUROSTAT (2016b). UOE Data Collection on Education Systems, Volume 2, Questionnaires and Instructions for their completion and submission, UNESCO Institute of Statistics, Montreal, OECD, Paris, and Eurostat, Luxembourg.

UNESCO Institute for Lifelong Learning (UIL) (2015a). Global Inventory of Regional and National Qualifications Frameworks, Volume I: Thematic Chapters. Hamburg: UIL., European Training Foundation (ETF), and the European Centre for the Development of Vocational Training (Cedefop).

UIL(2015b). Global Inventory of Regional and National Qualifications Frameworks, Volume II: National and Regional Cases. Hamburg: UIL., European Training Foundation (ETF), and the European Centre for the Development of Vocational Training (Cedefop).

UIS (2016a). Sustainable Development Data Digest: Laying the Foundation to Measure Sustainable Development Goal 4. Montreal: UNESCO. Retrieved from: http://tcg.uis.unesco.org/files/resources/laying-the-foundation-to-measure-sdg4-sustainable-development-data-digest-2016-en.pdf.

UIS (2016b). UIS Catalogue of Learning Assessments, Montreal, QB, UNESCO Institute for Statistics. Available at: www.uis.unesco.org/nada/en/index.php/catalogue/learning_assessments

UIS (2017a). The Global Alliance to Monitor Learning (GAML): Concept paper. Montreal: UIS. Available at http://uis.unesco.org/sites/default/files/documents/gaml-concept_paper-2017-en2_0.pdf.

UIS (2017b). The Global Alliance to Monitor Learning (GAML): Result Framework. Montreal: UIS. Available at: http://uis.unesco.org/sites/default/files/documents/gaml-result-framework-2017-en.pdf.

UIS (2017c). The Global Alliance to Monitor Learning Theory of Change. Montreal: UIS. Available at: http://uis.unesco.org/sites/default/files/documents/gaml-result-framework-2017-en.pdf.

UIS (2017d). SDG Data Reporting: Proposal of a Protocol for reporting Indicator 4.1.1. Montreal: UIS.

UIS (2017e). Constructing UIS proficiency scales and linking to assessments to support SDG Indicator 4.1.1 reporting. Discussion paper prepared by Management Systems International (MSI) for the 4th meeting of GAML. Montreal: UNESCO Institute of Statistics.

UIS (2017f). Mind the Gap: Proposal for a Standardised Measure for SDG 4 – Education 2030 Agenda. UIS Information Paper No. 46. Montreal: UIS. Available at: http://uis.unesco.org/sites/default/files/documents/unesco-infopaper-sdg_data_gaps-01.pdf.

Umar, Z. (2017, October 11). Constructing the 2017 Ibrahim Index of African Governance. Retrieved from: http://mo.ibrahim.foundation/news/2017/constructing-2017-ibrahim-index-african-governance/.

UNICEF (2017). Annual Results Report 2016: Education. New York: United Nations Children's Fund. Available at: https://www.unicef.org/publicpartnerships/files/2016arr_education.pdf.

UN Open Working Group on Sustainable Development Goals (2014). Outcome Document. New York. (UN Doc. A/68/970). Available at: http://sustainabledevelopment.un.org/focussdgs.html

Van Bergeijk, P.A.G., & van der Hoeven, R., eds. (2017). Sustainable Development Goals and Income Inequality. Cheltenham: Edward Elgar Publishing.

Vandemoortele, J. and E. Delamonica. 2010. Taking the MDGs Beyond 2015: Hasten Slowly. *Poverty in Focus.* No. 19. Available at: http://www.ipc undp.org/pub/IPCPovertyInFocus19.pdf

Verger, A., Novelli, M., & Altinyelken, H.K. (2013). Global Education Policy and International Development: An Introductory Framework. In Verger, A., M. Novelli and H. K. Altinyelken (eds.), Global Education Policy and International Development: New Agendas, Issues and Policies. Continuum: London.

World Bank. 2018. World Development Report 2018: Learning to Realize Education's Promise. Washington, DC: World Bank.

Wulff, A. (2017). Cashing in on SDG 4. In B. Adams, R. Bissio, C. Y. Ling, K. Donald, J. Martens, S. Stefano, & S. Vermuyten, *Spotlight on Sustainable Development 2017: Reclaiming policies for the public* (pp. 57-63). Beirut: The Reflection Group on the 2030 Agenda for Sustainable Development.

Appendix:

Sustainable Development Goal 4 and Global Indicator Framework *(in italics)*
Goal: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

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

4.2.2 Participation rate in organized learning (one year before the official primary entry age), by sex

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

4.3.1 Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, 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

4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill

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

4.5.1 Parity indices (female/male, rural/urban, bottom/top wealth quintile and others such as disability status, indigenous peoples and conflict-affected, as data become available) for all education indicators on this list that can be disaggregated

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

4.7.1 Extent to which (i) global citizenship education and (ii) education for sustainable development, including gender equality and human rights, are mainstreamed at all levels in: (a) national education policies, (b) curricula, (c) teacher education and (d) student assessment

4.a By 2030, build and upgrade education facilities that are child, disability and gender sensitive and provide safe, non-violent, inclusive and effective learning environments for all

4.a.1 Proportion of schools with access to: (a) electricity; (b) the Internet for pedagogical purposes; (c) computers for pedagogical purposes; (d) adapted infrastructure and materials for students with disabilities; (e) basic drinking water; (f) single-sex basic sanitation facilities; and (g) basic handwashing facilities (as per the WASH indicator definitions)

4.b By 2020, substantially expand globally the number of scholarships available to developing countries, in particular least developed countries, small island developing States and African countries, for enrolment in higher education, including vocational training and information and communications technology, technical, engineering and scientific programmes, in developed countries and other developing countries

4.b.1 Volume of official development assistance flows for scholarships by sector and type of study

4.c By 2030, substantially increase the supply of qualified teachers, including through international cooperation for teacher training in developing countries, especially least developed countries and small island developing States

4.c.1 Proportion of teachers in: (a) pre-primary; (b) primary; (c) lower secondary; and (d) upper secondary education who have received at least the minimum organized teacher training (e.g. pedagogical training) pre-service or in-service required for teaching at the relevant level in a given country