United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

SUSTAINABLE DEVELOPMENT GOALS

Information Paper No. 53
January 2019
UIS/2019/ED/IP/53

Costs and Benefits of Different Approaches to Measuring the Learning Proficiency of Students (SDG Indicator 4.1.1)

UNESCO

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

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

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

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication.

The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

This paper was written by Martin Gustafsson, Research on Socio-Economic Policy (ReSEP), Department of Economics, University of Stellenbosch

Published in 2019 by:

UNESCO Institute for Statistics
P.O. Box 6128, Succursale Centre-Ville
Montreal, Quebec H3C 3J7
Canada

Tel: +1 514-343-6880
Email: uis.publications@unesco.org
http://www.uis.unesco.org

Ref: UIS/2019/ED/IP53

Summary

The purpose of the current report is to advance policy debates around how countries and international organizations such as the UNESCO Institute for Statistics (UIS) measure learning outcomes and define proficiency levels in order to achieve the Sustainable Development Goal for education (SDG 4). This paper synthesises the recent debates and highlights issues which may not have received sufficient attention to arrive at informed proposals and recommendations. The report focusses on Indicator 4.1.1, which deals with children's proficiency in reading and mathematics across three education levels.

Section 2 discusses a range of important background issues. One important argument is that the comparability of national statistics over time should receive more attention. Up until now, much of the focus has been on the comparability of proficiency statistics across assessment programmes and countries. The latter is important, but focussing on the comparability of national statistics over time is vital in terms of UNESCO's commitment to global progress, and implies a somewhat different strategy to those associated with improving comparability across countries. One can think of good comparability in statistics over time, combined with a relatively crude degree of cross-country comparability as a second-best option which can still guide global strategies in powerful ways.

It is, moreover, argued in Section 2 that achieving statistics which are comparable across programmes and countries is perhaps even more difficult than is often assumed. One reason for this is that different parts of the world have different traditions when it comes to the stringency of proficiency benchmarks at different grades. Some parts of the world appear to apply more stringent benchmarks at lower primary grades relative to upper primary grades, while elsewhere the reverse applies. This obviously makes it more difficult to reach a global consensus around proficiency benchmarks. Moreover, these realities further complicate comparisons across countries, which often involve comparing slightly different grades, even at the same education level.

It is further argued that building advanced human capacity in the area of assessments across all countries is vital. Much of the required innovation is country-specific, meaning countries need local expertise.

Section 3 provides a new account of the presence of different assessment programmes around the world, designed to answer questions specific to the current report. This account draws distinctions not typically made in these kinds of analyses. For instance, in considering national assessment systems, it rates sample-based systems as generally being better sources of national trend statistics than censal systems. Moreover, the account provided here does not exclude national examinations, as these are conceivably a useful source of data, even though they are inherently limited. The account confirms that including national, and not just cross-national, programmes in the Indicator 4.1.1 reporting systems is necessary, at least until more countries participate in cross-national programmes. Without drawing from national programmes, the coverage of countries would be unacceptably low.

Section 4 describes three proposals which have been put forward for reporting on Indicator 4.1.1. The proposals are clearly not mutually exclusive, and one of the proposals described here is in fact a combination of two separately developed but similar proposals. A proposal labelled statistical recalibration of existing data involves using adjusted statistics from the cross-national programmes, where adjustments take advantage of the fact that some countries participate in more than one programme. The second proposal, labelled pedagogically informed determination of cut scores or social moderation, involves investment in a

global reporting scale, which would describe competencies that should be acquired at different points in the schooling system, and could then be used by countries to locate proficiency cut scores in national programmes which would be roughly comparable to those from other countries. The third proposal, recalibration through the running of parallel tests, or the Rosetta Stone approach, involves running new tests for sub-samples of students in existing programmes in order to establish equivalent cut scores from different programmes.

Section 5 provides a somewhat formal analysis of the costs and benefits of the three proposals, by considering the types of assessments each proposal draws from, and the benefits each of the assessment types bring. The first and third proposals are limited insofar as they draw only from cross-national programmes. Yet the first proposal is clearly beneficial insofar as it provides unique insights into differences across countries and over time for a large portion of the world. Moreover, it does so at little additional cost. The second social moderation proposal, which facilitates the use of data from national assessments and hence expands coverage significantly, also creates important opportunities for building capacity in the area of assessments in a large range of countries.

Section 6 describes a possible way forward. In part, this description is meant to illustrate the complexity of measuring, education data, and the practicalities of gathering and presenting national statistics on proficiency. Clearly, many aspects of the future scenario presented here are debatable. In this scenario, a version of the social moderation proposal would constitute the predominant approach. At the core would be the reporting scale referred to above. This would be the basis for producing roughly comparable proficiency statistics based on cross-national programmes. New UIS questionnaires, inserted within the existing set of questionnaires that have been used for many years by the UIS, would collect statistics and background technical information relating to national assessment programmes, both sample-based and censal. Crucially, these new questionnaires would specify that the UIS reporting scale would be used to determine cut scores within the national programmes. The UIS would maintain separate tables for statistics derived from cross-national programmes, and statistics derived from national programmes. However, where it is necessary to choose between the two, for instance in UNESCO's Global Education Monitoring Report, cross-national programme data would be considered preferable to national programme data.

Table of contents

List of tables

List of figures

1 Purpose of this report

The UN's Sustainable Development Goal for education (SDG 4) has shifted the global education agenda decisively towards learning outcomes and the proficiency levels attained by school children around the world. The UN's recent adoption of indicators focussing on the attainment of specific proficiency levels raises exciting and complex questions on how the UNESCO Institute for Statistics (UIS) and other stakeholders will move forward in measuring and reporting learning. The approach promoted by the UIS will have far-reaching implications not just for the quality and relevance of international statistics but also for how over 200 national education authorities measure learning and improve access to quality education.

Much has already been written about optimal approaches, and the factors that should influence the choices. It is now abundantly clear that determining a global data collection strategy is a technically complex matter, with serious cost and behavioural implications at various levels. It is a strategy that could easily be contested. As one report points out:[1]

> Both the political agendas and monitoring frameworks of the SDGs and Education 2030 are extremely ambitious. They demand an unprecedented increase in the collection, processing and dissemination from and, most importantly, within countries.

Political commitments and investments have already been made with respect to a broader strategy. This obviously influences what choices can be made in the future.

The focus here (as in several previous reports) is on Indicator 4.1.1, which aims to measure learning outcomes at three levels of schooling: lower primary, upper primary, and lower secondary, and for two subject areas: mathematics and reading. The indicator is defined as follows:[2]

> *4.1.1 Proportion of children and young people: (a) in Grade 2 or 3; (b) at the end of primary education; and (c) at the end of lower secondary education achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex*

Indicator 4.1.1 is one of 11 global education indicators enjoying a particularly high priority within the Education 2030 Agenda. There are also 32 thematic education indicators, which should also be reported by each country.[3]

The UIS has already published Indicator 4.1.1 proficiency statistics on its online database UIS.Stat[4], for years 2000 to 2015. To illustrate, 97 of 224 countries have at least one reading value for either of the two primary levels (a) and (b). These values are derived from cross-national assessment programmes, and use proficiency benchmarks developed separately in each of the programmes, even though they were not intended to be

[1] UIS, 2017h, p. 2.
[2] United Nations, 2017.
[3] UIS, 2017k.
[4] http://data.uis.unesco.org (accessed June 2018).

comparable across programmes.[5] This approach is provisional in the absence of a more comprehensive and country-driven system.

Approaches that have been put forward differ most obviously in terms of their technical complexity, financial cost, and implied comparability of national statistics. Less obvious differences relate to their sustainability over time, their impact on the politics, planning and operations of national education authorities, their ability to contribute to capacity building within countries, and their persuasive power in the media and policy debates. There are several ways in which existing proposals could be taken forward. Hybrid approaches are possible. For instance, the prioritisation of new data collection programmes could be staggered according to the three levels of the schooling system covered by Indicator 4.1.1. Initially it may be best to prioritise the measurement of learning outcomes at the primary level, for example. It is likely that any strategy will have to accept migration over time from weaker to better data collection systems. In other words, an all-or-nothing approach of relying only on the ideal is not possible.

How is the current report different from what has been produced previously? This report evaluates, in broad and not necessarily monetary terms, the costs and benefits of different approaches to measuring Indicator 4.1.1. An earlier UIS report has to some extent examined the costs and benefits of reporting all 43 global and thematic SDG 4 indicators[6]. The current report deals with just one indicator, but an indicator which is new, has large cost implications, and is particularly complex. Another earlier report briefly evaluated different strategies for advancing the measurement of Indicator 4.1.1, but did not do so in much depth, or within a cost-benefit framework[7]. It should be emphasised that although the current report considers relative costs carefully, and makes reference to a few financial costs, its intention is not to provide actual proposed budgets for the various proposals.

Section 2 discusses several critical issues that influence the analysis of the costs and benefits. These include how one assesses costs and benefits in education, the capacity required within countries to deal with new reporting systems on learning outcomes, the direct financial costs of assessments, the comparability of results across countries, and over time, the advantages and disadvantages of particular types of assessments, the problem of out-of-school children, the time it takes assessment systems to produce results, the dynamics of education policy debates within countries and ways in which the UIS incentivises effective data collection from countries. The aim of this section is not so much to repeat what several other reports have already explained, but to highlight issues that may not have been clear enough previously, and which influence the cost-benefit analysis of the current report.

Section 3 provides a new audit of assessment systems of potential use for Indicator 4.1.1. These kinds of audits have been undertaken before, but it seemed a new one was necessary that was geared for the specific concerns of the current report. The audit presented here pays special attention to the distribution of a specific type of assessment.

[5] Altinok (2017, pp. 13-14) summarises these programme-specific benchmarks. Where choices could be made within a programme on what benchmark to use, it appears the benchmarks put forward in Altinok were used.
[6] UIS, 2017h.
[7] UIS, 2017d.

Section 4 describes three existing proposals for advancing reporting on Indicator 4.1.1. The proposals are:

1. Statistical recalibration of existing data. This is currently based largely on work done by Nadir Altinok.

2. Pedagogically informed determination of cut scores (social moderation). This draws from proposals produced by the Australian Council for Educational Research (ACER, 2017) and Management Systems International (MSI). The proposals of the two organizations differ in important respects, but they are treated as one proposal here due to their overlaps and the benefits of mixing the two proposals. The term social moderation is used just by MSI.

3. Recalibration through the running of parallel tests (Rosetta Stone). This draws from a proposal by the IEA.[8]

These proposals are the point of departure for the cost-benefit analysis, though the possibility of combining and adjusting them also receives attention. This section also highlights elements from the three proposals to which the UIS has already committed itself.

Section 5 provides the actual cost-benefit analysis. A framework for the analysis is presented, followed by conclusions relating to the three proposals introduced previously. It is not the intention of the current report to provide a final recommendation on which of the three approaches (or what combination of the three) to follow. Rather, the intention is to facilitate final decisions.

Section 6 concludes the report by describing a few ways forward for Indicator 4.1.1 which seem feasible and possibly optimal, and which draw from the analysis of the preceding sections. It is by no means a widely consulted plan, but rather an input into the planning processes currently occurring.

2 Critical issues

2.1 Understanding costs and benefits in education planning

Key points: In this report's informal evaluation of costs and benefits associated with different Indicator 4.1.1 approaches, certain economic concepts seem worth bearing in mind: positive externalities, opportunity costs, the centrality of human capital, balancing imitation and innovation, and the apparent trade-off between efficiency and equity.

The current report uses concepts from economics and cost-benefit analysis[9] in approaching the question of optimal approaches for reporting Indicator 4.1.1. It is by no means a formal cost-benefit analysis. In fact, formal cost-benefit analysis is rare in education planning, due to its complexity and data demands.

So, what are some of the key concepts worth keeping in mind? With respect to benefits, the UIS and its partners are not just interested in collecting statistics for their own sake, but in establishing a data collection system, and refining existing ones, in a manner whereby (a) the very process of collecting data has positive side-effects and (b) the statistics are used constructively to bring about better education policies that

[8] International Association for the Evaluation of Educational Achievement.
[9] See Woodhall (2004) and Cheng et al (2002).

advance the SDGs. In economics, beneficial side-effects are referred to as positive externalities. People tend to value benefits obtained sooner, rather than later. We thus discount to some extent benefits expected in the more distant future. It is good to make this as explicit as possible in the analysis. There may be a future ideal, perhaps involving high costs in the near future, which one is prepared to lose because the ideal would be realised so far into the future. In the area of assessments, building a single cross-national testing system covering all countries could be such an ideal that one agrees not to pursue.

Turning to costs, direct budgetary costs for organizations and countries are influential. Such costs may appear high enough to dissuade governments from investing in assessment systems, particularly if faced with resistance from voters or teacher unions. This is more likely to happen if the benefits of these investments are not clearly communicated.

The concept of opportunity costs is important. This is the loss resulting from spending money on a relatively inefficient programme, when a more efficient alternative would produce better results. In planning, it can be especially important to point out the opportunity costs of well-established programmes which are assumed to be worthwhile, but are perhaps not when one considers the benefits of utilising the programme's budget on something else. Put differently, a careful consideration of the opportunity costs of existing programmes can help to bring about, and finance, innovation. Opportunity costs arise not only because budgets are devoted to the wrong things but because even in a policy discourse, too much attention paid to more traditional issues can crowd out discussions and thinking relating to innovations.

Any programme requires human capital. It should not be difficult to convince education planners of this. Yet, as will be argued below, the importance of advanced human capital in the area of assessments has probably been under-estimated. Effective tools and guidelines represent essential technological capital, but their utility may be reduced if there are not enough specialists with a deeper knowledge of assessment issues across all countries. In development economics, countries move forward because they employ the right mix of good imitation and innovation. This can be applied to assessments. To illustrate, to some extent it is necessary to have clear instructions which non-psychometricians can follow in order to rescale test scores – this is replication, or good imitation. However, it is also necessary to have enough expert psychometricians in each country, as such people bring about country-specific innovations, are likely to facilitate better implementation of existing instructions and can provide criticism and feedback which can help to improve the instructions.

Finally, there is often a trade-off between efficiency and equity in education planning. In the Indicator 4.1.1 context, there is a clear and justified need to compile international sets of statistics which are as accurate as possible. The most efficient and least time-consuming way to go about this may be to place a strong emphasis on the large global agencies that assess students across the world. Yet, this could undermine equity, which is a central concern of the entire SDG framework. Advancing equity in education is in part a question of reducing proficiency inequalities across populations, but it is also about empowering governments and education planners in developing countries. The costs associated with extensive capacity building may make more country-driven approaches less efficient in a narrow sense, but this type of investment carries important equity benefits and enhances buy-in and ownership of the monitoring system among countries.

2.2 Understanding the country-level capacity building required

Key points: International education statistics would be in a healthier state if the utility of (or demand for) statistics were taken into account in better ways. There has probably not been sufficient emphasis on building advanced human capacity in assessment-related areas across all countries. There is a need for better technical documentation to guide countries in the process of generating proficiency statistics. This documentation should include accounts of actual country experiences.

The way data is used in education planning and management has become increasingly global with the expansion of organizations such as the UIS and easier access to knowledge through the internet. Bringing together data and knowledge about learning outcomes from around the world will strengthen this trend. The trend creates opportunities for innovation by international organizations or individual countries. Interesting innovations in one place are likely to lead to change elsewhere. One could argue that this has created an enabling environment for the 'data revolution' referred to in various UN and UNESCO proposals.[10]

Especially in education, this data revolution seems overdue. As one UIS report points out:[11]

> Evidence from the health sector strongly suggests that interest in data preceded, and led to, the hugely disproportionate investment made in data systems in that sector relative to the education sector.

Another UIS report suggests that the investment needed is not so much additional funds – funding is already fairly high at an aggregate level – but rather more innovative use of existing financial and human resources.[12]

So what investments in what kinds of innovations are needed? For the purposes of the current report, what stands out is the need for data work that focusses more on making a difference to the sector. In other words, a stronger emphasis is needed on the demand and utilisation of data, not simply supplying data.[13] This requires thinking differently and more broadly about processes around data. For this, human capacity is needed, both with respect to broad strategic thinking around data, but also with respect to very specific skills which will be discussed below.

It is worth noting that human capacity appears under-emphasised in the current literature on education data. In particular, human capacity to bring about innovation within individual countries seems under-emphasised. Instead, much of the emphasis falls on tools in the form of manuals and standards. These tools are important, but on their own, are not a guarantee that the necessary human capacity will be built. The cross-national assessment programmes have created networks that have facilitated country-specific capacity building, yet the processes within these programmes are to a large degree premised on a model where innovation and advanced technical work, for instance with respect to sampling and psychometrics, occurs in one place, while each country follows a set of instructions. The problem with insufficient innovation (as opposed to imitation) in individual countries is that innovative and country-focussed use of the data

[10] UIS, 2017k; United Nations: Data Revolution Group, 2014.
[11] UIS, 2017h, p. 9.
[12] UIS, 2018[a], p. 6.
[13] UIS, 2018a.

emerging from the cross-national programme is often limited, and the capacity to design national programmes is limited. Moreover, weak technical capacity in a country means that national assessment systems succumb more easily to political interference, a real risk in an area such as assessments.

One concrete way in which these problems can be addressed, according to the Global Alliance for Monitoring Learning (GAML), an initiative linked to the UIS, is to clarify and strengthen leadership:[14]

> Currently, no global player is taking the lead in the provision of capacity-building services for learning assessments.

A good practice guide provides a basic list of assessment-related skills which can be considered advanced, and which, it is argued, should perhaps be secured through outsourcing:[15]

> … item writing, translation, linguistic quality assurance, graphic design, sampling, data processing, psychometric analysis, editing or publishing.

To this list can be added skills relating to the dissemination of data, such as skills in developing technical documentation accompanying the data, or metadata, and skills needed to anonymise data. It could be argued that any country or education authority should aim to have these competencies within the authority, or at least the country. In other words, the aim should be to reduce the need for outsourcing. Though advanced, these skills can be considered essential for sustaining and defending an effective national assessment system. What would probably be beneficial for capacity building is an elaborated version of the above list of competencies, to assist in particular developing countries to identify what skills should be developed.

Even with respect to the tools required for assessments, innovation is necessary. Above all, investments in guides for designing national assessments with sufficient technical depth and a sufficient grounding in the actual experiences of countries seems needed. To illustrate, Chile has a relatively advanced national assessment system in part because its design draws from Standards for Educational and Psychological Testing, a manual developed by three psychological and testing organizations in the United States, and used extensively in the United States.[16] However, for many government officials and academics in developing countries wishing to improve their assessment-related competencies, this guide is prohibitively expensive and may not be suitable across a large range of countries. Yet, it represents the type of technical guidance which should become increasingly available to education assessment specialists around the world.

[14] UIS, 2017i.
[15] UIS, 2017j, p. 14.
[16] Chile: Agencia de la Calidad de la Educación, 2014: 90.

2.3 Direct financial costs of assessments for countries

Key points: Even for developing countries, the cost of assessing outcomes systematically is low relative to the overall cost of providing schooling. Assessment systems, if well-designed, can have positive impacts that go beyond simply producing statistics.

Assessments required to report SDG 4 indicators are relatively costly compared to other data collection systems required for these indicators. It is estimated that data on the quality of learning, or proficiency levels, will account for around a half of all costs related to SDG 4 reporting.[17]

However, relative to the overall cost of providing schooling, assessment systems appear not to be costly. A World Bank report estimates that assessments should not account for more than around 0.3% of overall public spending on schooling[18]. Given the value of having good information on learning outcomes, such a figure can be considered negligible. Existing estimates of the cost of running a sample-based assessment in a country range from US$500,000 to around US$1.7 million.[19] This is the cost of one cycle of an assessment programme. One can expect costs in initial cycles to be higher than in subsequent cycles due to the need for start-up and development activities.

Participation in one round of a large international assessment programme such as TIMSS[20] and PISA[21] costs a country around US$800,000.[22] The figure is a lower, US$200,000 to US$500,000, for regional cross-national programmes such as LLECE[23] and PASEC.[24]

Given that the costs of a sample-based assessment, as well as the optimal sample size, are largely independent of the size of the country, the ratio of assessment costs to overall spending becomes higher in smaller countries. Some basic number checking: assuming a cost of US$750,000 per assessment that serves three grades (this is roughly the requirement of Indicator 4.1.1), confirms that the ratio for many developing countries indeed comes to around the 0.3% estimated by the World Bank. However, small countries, or countries with low spending per student, reach a percentage of 1.0% Mauritius, Malawi and Swaziland would all fall into this category. A very small country such as Cabo Verde would see the cost of assessments coming to 3.0% of overall spending. For this reason, it has been argued that for some countries, running both a national programme and participating in a cross-national one at the same level of schooling might not be cost-effective.[25]

---

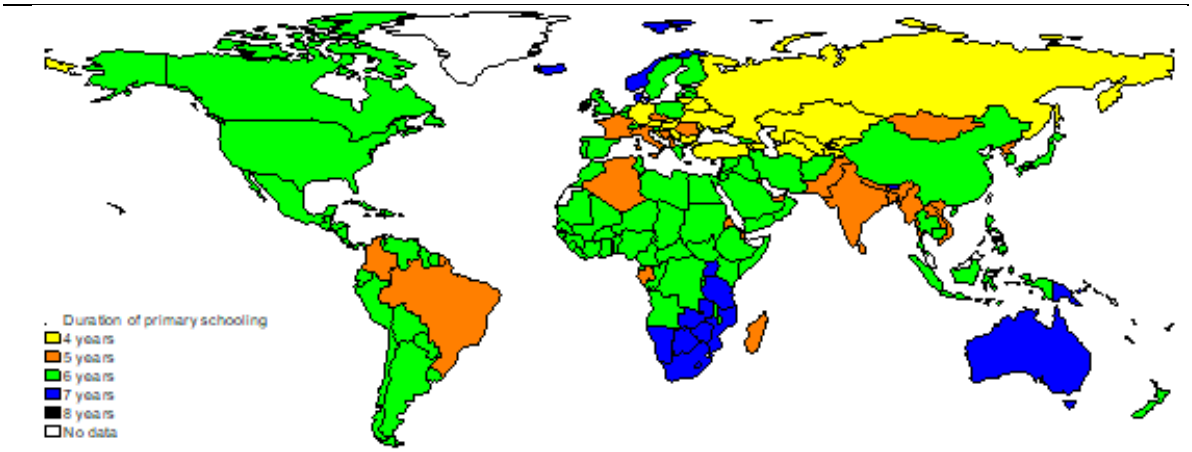
[17] UIS, 2017h, p. 27.
[18] Clarke, 2012, p. 6.
[19] UIS, 2017h: 26; Brookings Institute, 2015, p. 9.
[20] Trends in International Mathematics and Science Study.
[21] Programme for International Student Assessment.
[22] UIS, 2018b, pp. 19-20.
[23] Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education).
[24] Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (Analysis Programme of the CONFEMEN Education Systems).
[25] UIS, 2018b, p. 17.

2.4 The comparability of grades and education levels

Key points: The fact that primary schooling has a different duration in different countries means a term such as 'the end of primary' can mean different things in different places. The fact that the gap between proficiency benchmarks and reality tends to be systematically correlated to grade level within countries and regions complicates comparisons across countries and assessment programmes, in particular where the grade is not identical.

The challenge posed by the fact that different schooling systems and different cross-national assessment programmes conduct assessments at different grades, and the fact that the relationship between age and grade differs by country, has been discussed previously.[26] These realities mean that Indicator 4.1.1 values for any one of the three education levels will in many cases be comparing rather different groups of children, depending on the country or assessment programme. Part of the problem is illustrated by **Figure 1**, which shows that while six years of primary schooling is most common across countries, there are regionally concentrated differences, such as seven years in Southern Africa, five years in much of South Asia, and four years in many Central Asian countries. This would influence the ages and grades of children considered by countries to be at the end of primary, and even the meaning of lower secondary.

Figure 1. Duration of primary schooling



Source: UIS.Stat. Data refer to 2017.

One issue appears not to have received attention in previous reports: the fact that proficiency statistics vary within a country by grade in an apparently systematic manner, and that these patterns are repeated across many countries. Recent literature on gaps between curriculum expectations and actual learning[27] suggests that as the grade increases, proficiency levels drop as expectations, relative to reality, rise. However, as will be seen below, the situation is less straightforward.

[26] UIS, 2017d, p. 9.
[27] Pritchett and Beatty, 2012.

Figure 2 displays the proficiency statistics for 2013 and for reading from UIS.Stat, for the levels lower primary and end of primary. All the points in the graph refer to Latin American countries participating in LLECE as only this programme produced statistics for these education levels and for 2013. For all countries, fewer children were proficient at the lower primary level than at the end of primary. This suggests that relative to actual learning outcomes, expectations were higher for Grade 3 than for Grade 6 (these are the two grades covered by LLECE). A similar pattern, though it is less clear, emerges if one performs the same analysis using 2014 UIS.Stat data *(see **Figure 3**).* That analysis reflects just Francophone African countries participating in PASEC, which tested Grades 2 and 5.

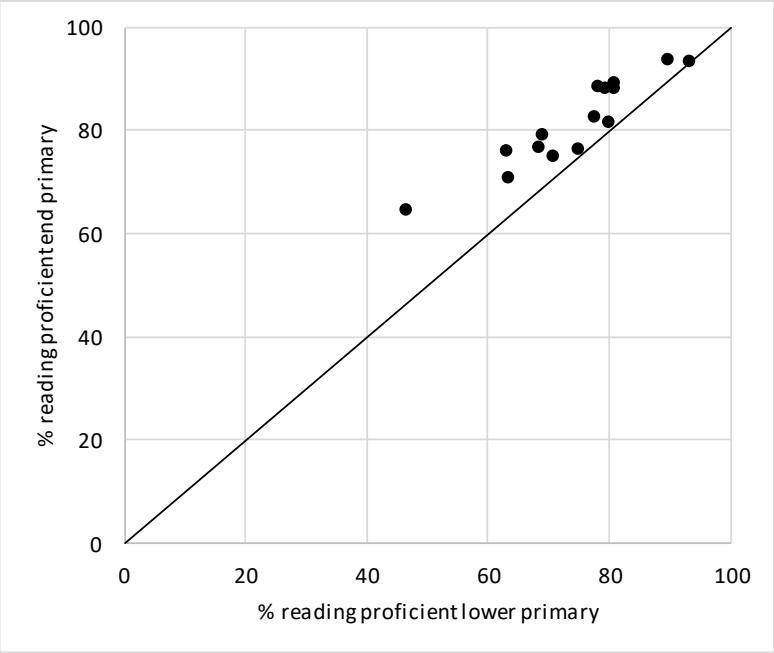
At a national level, South Africa's national assessment system has displayed the reverse of what one sees in the two graphs, namely, far higher levels of proficiency at lower grades in a system that tested Grades 1 to 6 and Grade 9.[28] Ghana's national assessment, on the other hand, has shown patterns more consistent with what is seen in Figures 1 to 3.[29] The point is that if countries and whole regions tend systematically to use standards which are either more or less stringent the higher the grade, then this contributes to the non-comparability across countries of proficiency statistics, specifically where one compares different grades, for instance, Grades 5 and 6 (a comparison one might easily make if one wanted to compare LLECE and PASEC countries).

Grade repetition as reported to the UIS can be used to provide insights into how different countries have historically applied different proficiency standards across grades.

[28] South Africa: Department of Basic Education, 2016, p. 35.
[29] Ghana: Ministry of Education, 2014.

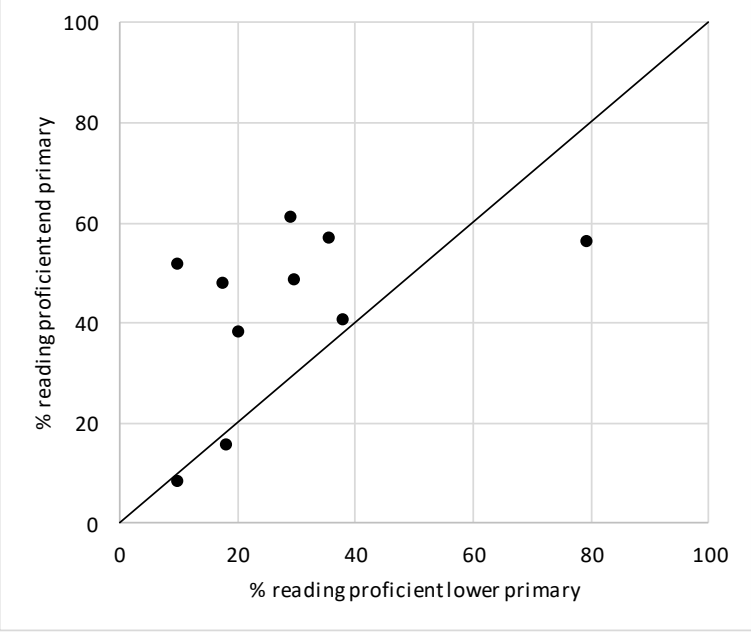
Figure *4* uses 2011 values for the indicator "percentage of repeaters in Grade X of primary education" on UIS.Stat. Countries from a variety of regions with relatively striking patterns were chosen. Malawi and Cambodia display fairly consistent patterns whereby grade repetition becomes lower the higher the grade. There could be several factors behind this, including the dropping out of weaker students in earlier grades. However, one factor is likely to be teachers' higher expectations, relative to the actual competencies of students, at the lower grades. Côte d'Ivoire displays a very different pattern, with around 15% of students repeating in Grades 1 to 4, and then a spike of 50% in Grade 5. Côte d'Ivoire has an examination in Grade 6 which could result in entry restrictions into this grade, hence high repetition in the previous grade.

Figure 2. Proficiency in lower primary against end of primary education, LLECE



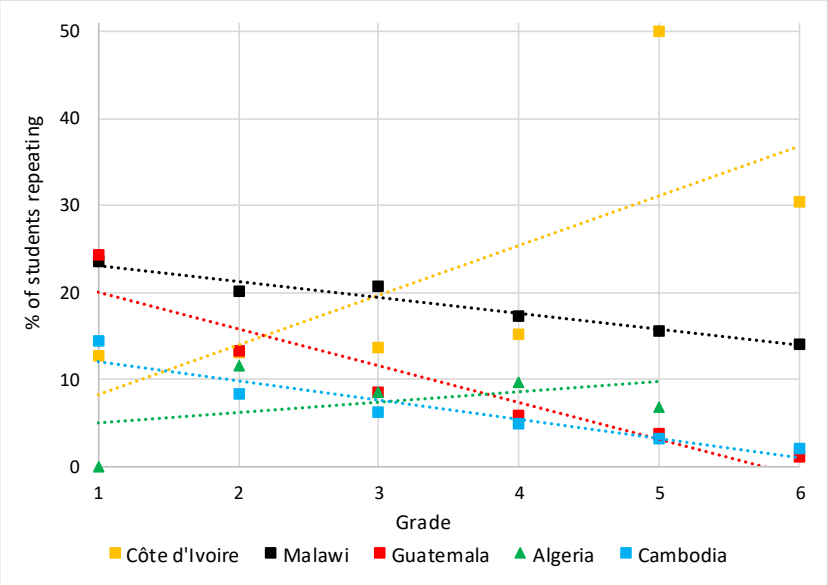
Source: LLECE, 2013.

Figure 3. Proficiency in lower primary against end of primary education, PASEC



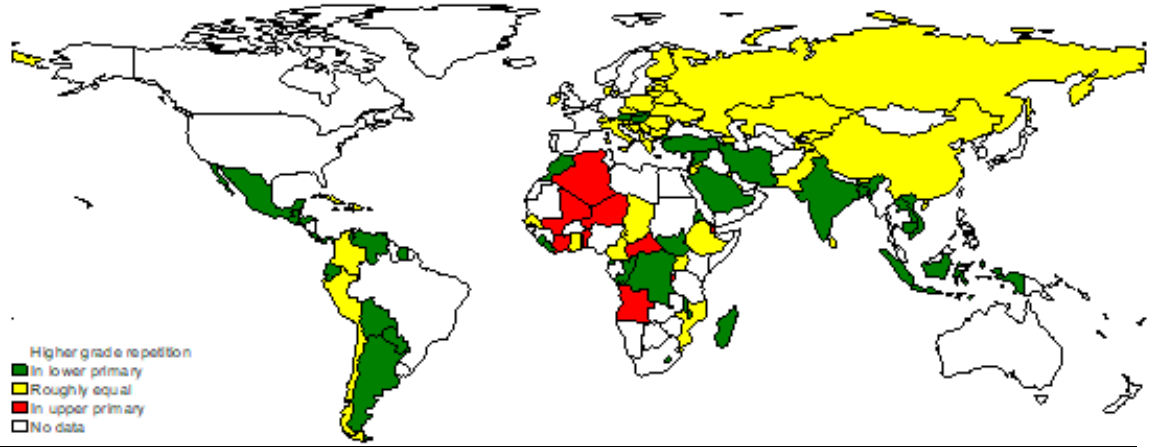
Source: PASEC, 2014.

Figure 4. Grade repetition by grade and country



Is there a geographical pattern with respect to the slopes seen in Figure 4? **Figure 5** suggests there is. Developing countries are more likely to display a non-constant pattern, with either more repetition at the lower primary level, or more at the upper primary level. However, few countries display more repetition at the upper primary level, and these are largely a group of Francophone sub-Saharan African countries.

Figure 5. Patterns of grade repetition at the primary level



Source: UIS.Stat. Values from 2011 used. Slopes as shown in ***Figure 4*** *were calculated. Any country with a grade-on-grade slope greater than 0.5 percentage points was considered to have more repetition at the upper primary level. A value less than -0.5 was considered to indicate more repetition at the lower primary level.*

2.5 Comparability of assessment results across space and time

Key points: The importance of comparability of national results over time and how to achieve this, has probably not received enough attention. While the comparability of statistics across countries influences comparability over time, to some extent the latter operates independently of the former. Even the relatively well-staffed cross-national assessment programmes have displayed inconsistent trend data. Strategies are needed to minimise this.

What is emphasised strongly in many reports relating to Indicator 4.1.1 is the need to produce internationally comparable statistics on learning outcomes. Most of these reports acknowledge that this is not easy to achieve, for various reasons. Countries may wish to use just nationally determined proficiency benchmarks which are meaningful to the country. Even if there is the political will to adopt global proficiency benchmarks, the fragmented nature of the current landscape of cross-national and national assessment systems would make the goal of internationally comparable statistics difficult to achieve.

It is useful to examine in some detail the assumption, implied by much of the existing discussion around measuring Indicator 4.1.1, that more internationally comparable statistics on learning outcomes would contribute to better quality schooling around the world.

We do undoubtedly need to measure change over time with respect to learning outcomes and the attainment of proficiency benchmarks. If we do not do this, we will not know whether we are making progress towards SDG 4, and this in turn would make it very difficult to determine whether strategies adopted around the world to reach the goal were delivering the desired results. But how does improving the comparability of statistics across countries and assessment programmes, something which has been prioritised, help us gauge progress towards the achievement of 'relevant and effective learning outcomes' (as stipulated by SDG Target 4.1)[30] for all young people? The logic is simple. In making statistics on learning outcomes comparable across countries, and more specifically across assessment programmes, at one point in time, through some kind of equating or linking methodology, and assuming that each assessment programme produces statistics which are comparable over time, it will produce statistics even in future years which are comparable across countries. This will enable global aggregate statistics over time to reflect the degree of progress.

This logic can be said to reflect an ideal, an ideal where investing in better comparability across programmes and countries has good returns. There is a second-best approach, however, which has probably not been sufficiently appreciated. In the second-best approach, it is accepted that the comparability of statistics across countries will be somewhat limited, even after recalibrations, but considerable effort is put into improving the comparability of each learning assessment programme and each country's national statistics over time. In such an approach, the global aggregate statistics are somewhat crude, because the underlying national statistics are only roughly comparable to each other, but programme- and country-level statistics are able to provide relatively reliable trend data. Thus, if all countries, or virtually all countries, are displaying improvements over time, it is highly certain that the world as a whole is displaying improvements. The magnitude of the global improvements could be calculated in a crude sense, though not as accurately as in

[30] United Nations, 2017, p. 6.

the ideal measurement approach. But country-level magnitudes of improvement would be reliable and certainly meaningful and useful to the citizens and governments of individual countries.

A key argument of the current report is that debates around optimal monitoring strategies for learning outcomes should look beyond the comparability of national statistics across countries (across space), and focus more on the comparability of national statistics over time. It should be acknowledged that the two aspects, space and time, are interrelated, but also to some degree independent of each other.

Note it is not being argued that striving for comparability of statistics across countries is unimportant. This is important and efforts in this regard have been vital for improving our knowledge of learning and schooling. Rather, what is argued here is that the comparability of statistics, and why this is important, should be thought through more systematically. Put differently, and in line with the discussion in Section 2.2, data collection systems should aim to produce statistics that serve a particular practical purpose, in other words, statistics that respond to a valid demand, and this should be made explicit.

If concerns over the comparability of learning outcome statistics over time were to be given more emphasis, what are the issues that should receive special attention? This question will not be answered comprehensively here, but some pointers will be provided.

It is instructive to note that even in the world's most technically advanced cross-national learning assessment programmes, concerns have been raised around the comparability over time of national statistics. An important case in this regard is Brazil's PISA results. Brazil's PISA improvements between 2000 and 2009 in mathematics were exceptionally large compared to gains seen in other countries. On average, the annual improvement was around 6 PISA points, or 0.06 of a PISA standard deviation. By any standards, this represents particularly steep progress. This trend led to considerable interest in understanding what Brazil did to achieve this change.[31] However, it seems as if Brazil's PISA results over-estimated the gains to a considerable extent. As pointed out in Klein (2011) and Carnoy et al (2015), changes in the date on which the PISA tests were written resulted in biases whose overall effect was an over-estimation of Brazil's gains over time. One would expect tests run later in the school year to produce better results than tests run earlier in the school year. Even after controlling for these distortions, Brazil's mathematics improvements are noteworthy, and come to around 3 PISA points a year.[32] Distortions caused by shifts in the test date are said to have affected not just Brazil's PISA results, but even the PISA results of other Latin American countries.

Jerrim (2013) has argued that in the case of England, a substantial decline in PISA mathematics over the period 2000 to 2009 is not a true reflection of actual trends, and that the actual trend is likely to be no change, or an improvement. TIMSS Grade 8 mathematics data in fact displayed a large improvement over the same period. Again, changes in the testing date probably distorted the PISA trend, but in addition, there were important changes in the sampling process, which one can assume would not all be in line with PISA specifications. It is suggested that an underlying reason for all this was that the responsibility for administering PISA in England shifted from one institution to another. Distortions in the data, Jerrim argues, resulted in a distorted policy discourse. Specifically, as a dataset from the Organisation for Economic Cooperation and Development (OECD) , the PISA data carried considerable weight, meaning policy debates

[31] Bruns, Evans and Luque, 2012.
[32] Carnoy et al., 2015, p. 11.

were to some degree premised on the existence of a decline in educational quality, which in fact probably did not exist.

While PISA is particularly difficult for a country to administer because of its focus on an age cohort, and not a school grade, it is not difficult to imagine that similar country-specific sampling and test administration problems would arise in other assessment programmes. With regard to TIMSS, Jerrim (2013) raises concerns around the fact that in certain years in England close to half of the initially sampled schools had to be replaced due to the refusal by schools to participate. Such high replacement ratios are clearly worrying, given the likelihood that schools which refuse are distinct with respect to their teaching and learning characteristics. Jerrim suggests changes in the replacement ratios could have contributed towards the steep increase in England's TIMSS Grade 8 mathematics scores.

If the large global programmes PISA and TIMSS have experienced country-level distortions in the measurement of educational progress, is seems likely that one would find similar, and probably more serious, problems in the more recently established and less resource-endowed regional programmes. Gustafsson and Nuga Deliwe (2017) have pointed to certain problems with the comparability between the 2000 and 2007 results of SACMEQ,[33] resulting in part from the absence of a matrix sampling approach in the design of the tests, and weaknesses in the test administration processes in some countries, factors which appear to have led to some cheating, though not of such a magnitude that SACMEQ country rankings would change. Moreover, they raise concerns about the insufficient availability of technical documentation, in particular on the equating between the two years, which makes it difficult to verify whether the process of converting the raw data to final country averages was sufficiently rigorous. The need for this verification is made particularly important due to the fact that relationships between the raw classical test results and Rasch scores exist which cannot easily be explained.

LLECE has published a relatively detailed account of how comparability between its 2006 SERCE assessments and its 2013 TERCE assessments (both at the Grades 3 and 6 levels) was achieved.[34] However, so far there appears to have been little interrogation of the publicly available SERCE and TERCE datasets, outside of LLECE, to verify the internal consistency of the results. The PISA literature referred to above demonstrates how important such external interrogation of the data is for reassurances of the reliability of, above all, trend statistics, or for understanding the need for caution in interpreting the data, or the need for some adjustment.

PASEC demonstrates how important it is for sufficient technical documentation to be made publicly available. The PASEC website includes a comprehensive report on the PASEC 2014 results,[35] but no indication of whether the 2014 results are comparable to earlier results. It can probably be assumed that results are not comparable over time, but it would be good if this were made explicit to prevent inappropriate comparisons to earlier years. In fact, the publicly available statistics on pre-2014 PASEC results, found for instance on UIS.Stat, suggest strongly that one cannot draw conclusions about trends over time. To illustrate, reading proficiency statistics on UIS.Stat for Cameroon are 92% and 49% for 2006 and 2014.[36] The average

[33] Southern and Eastern Africa Consortium for Monitoring Educational Quality.
[34] UNESCO, 2016.
[35] CONFEMEN, 2015.
[36] UIS.Stat data (accessed July 2018).

percentage point decline for the seven PASEC countries with end primary results for 2006 and 2014 was 21. Clearly, the countries cannot have experienced an educational deterioration of this magnitude.

To conclude, challenges that deserve close attention if comparability over time is to be strengthened include a better focus on how cross-national programmes are implemented within individual countries. While the large international learning assessment programmes, PISA, TIMSS and PIRLS,[37] are generally good at making public technical documentation relating to processes controlled directly by the programme's central office, processes controlled by implementers within countries are not as well documented, which makes verification of those processes difficult and reduces the public accountability of these implementers. In the regional programmes the problem is more serious in the sense that here even processes controlled by the central office, such as test design and equating across years, are not always sufficiently documented.

2.6 A hierarchy of assessment types

Key points: For the purposes of each of the three education levels of Indicator 4.1.1, it seems useful to think of four types of learning assessments, each offering specific opportunities and challenges. The four are, from most to least technically robust: (a) the three large international programmes (PISA, TIMSS and PIRLS); (b) five regional cross-national assessments; (c) sample-based national assessments; (d) censal assessments; and e) national examinations.

To some extent, a hierarchy of assessment types has been put forward in the literature, from more to less reliable, where reliability can in part be thought of in terms of comparability across space (for instance countries) or time.

One thing one can probably take as a given, based in part on the analysis in Section 2.5, is that the large international programmes of the OECD and IEA (PISA, TIMSS and PIRLS) are in general more rigorous, and thus allow for better comparability, than regional programmes such as SACMEQ and LLECE. The large programmes simply have a longer history, are better funded, and have easier access to the scarce skills needed for the more technical aspects of the work.

Moreover, it is easy to conclude that the cross-national assessment programmes are better at measuring trends than national programmes, because the former can access more resources, and are less exposed to political interference by national governments. This is likely to be particularly true in developing countries. However, even in developed countries, cross-national programmes may provide considerably more reliable statistics than a national programme. This has been demonstrated in the case of Italy, where the national programme has been shown to be susceptible to cheating, which in turn has distorted comparisons within the country in undesirable ways.[38]

There has been a strong emphasis on promoting national assessments above national examinations as a source of data on national learning trends. Under the heading 'National assessments are indispensable for informing policy', the 2013-14 Global Monitoring Report of UNESCO says the following:[39]

[37] Progress in International Reading Literacy Study.
[38] Ferrer-Esteban, 2013.
[39] UNESCO, 2014, p. 90.

> *…policymakers often consider their public examination system as equivalent to a national assessment system, even though the two serve very different purposes. Public examination systems are used to promote students between levels of education (and so set standards and benchmarks according to places available); national assessments should be a diagnostic tool that can establish whether students achieve the learning standards expected in the curriculum by a particular age or grade, and how this achievement changes over time for subgroups of the population.*

A World Bank guide on designing national assessments provides a table distinguishing between the features of assessments and examinations.[40] However, the distinction between the two may not be that clear. National assessments which are censal, as opposed to sample-based, can serve purposes similar to those of examinations. In particular, both permit something that education authorities value considerably: performance statistics for all schools which can be used for accountability purposes. Education authorities often violate the assessment-examinations distinction by using examinations to gauge progress in the schooling system (this is essentially what UNESCO is warning against). Are education authorities always wrong if they do this? This is explored below. Perhaps the key distinction of relevance to the current report is that examinations never use common test items to facilitate comparison across years, while ideally assessments should do this. This suggests that assessments should be better at gauging trends over time than examinations.

One question which ought to be clearer in the available guides and manuals is whether one compromises on comparability, in particular comparability over time, if one has a national assessment which is censal, as opposed to sample-based. One should expect sample-based assessments to be better at measuring change over time for the simple reason that sample-based programmes cover fewer schools and are therefore less exposed to the risk that tests are leaked, which means future tests will not be able to repeat non-exposed items, and will therefore not be anchored to earlier tests. Put differently, security around the tests becomes a larger challenge if all students across all schools are tested. There are clearly countries which do succeed in implementing the required level security in a censal assessment. Nevertheless, greater caution in interpreting trends seen in censal assessments (as opposed to sample-based assessments) seems necessary. Differences between two developing countries which both run censal assessments, Brazil and Chile, are instructive.

Brazil's Prova Brasil programme, which has been described as exemplary even by developed country standards,[41] uses anchor items to facilitate comparability over time.[42] However, details on this, but also details on other issues such as security during the test administration process and the extent of non-participation of schools, could be clearer, which in turn could strengthen further the credibility of the data and published performance trends. Chile's SIMCE appears more developed in this regard. Details on, for instance, non-participation are provided, and test administration processes are not only better described, they are also particularly rigorous.[43] SIMCE also seems to have a more developed process for sharing not just school-level, but even student-level data in the case of approved researchers. Not only does this contribute towards research in general, it allows for external validation of the data and official trend statistics.

---

[40] Greaney and Kellaghan, 2008.
[41] Bruns, Evans and Luque, 2012, p. 7.
[42] Brazil: Ministério de Educação, 2015ª, p. 30.
[43] Chile: Agencia de la Calidad de la Educación, 2014, pp. 79, 88.

Clearly the label 'assessment' in a programme is no guarantee that standard linking procedures across years, using common items, are used. To illustrate, South Africa's Annual National Assessment programme, while in some respects an assessment – it did not produce reports for individual students – was not designed to include common items. This limitation was acknowledged officially, yet the programme was often used to make comparisons across years. In this sense, the programme functioned like an examination.[44]

Is examinations data always unsuitable for monitoring progress in education? In theory at least, there are strong incentives to standardise examination results over time as not doing so results in the unequal treatment of different cohorts of students. However, in practice rigorous standardisation is difficult without secure common items. But how serious are the across-time comparability problems in examinations data? It is difficult to answer this question as little research has occurred in this area. Many countries do clearly attach much importance to examination-based trends, and use these trends for monitoring purposes. In particular in developing countries, examinations exist at the primary level, not just the secondary level. To give one example, Sri Lanka's Grade 5 Scholarship Examination, which has existed since the 1940s, is used to determine which students should fall within the roughly 10% who qualify for a more elite education. But is also determines whether students pass a more basic threshold – in recent years around 70% have.[45] Conceivably, there are similarities between this threshold and the proficiency benchmarks envisaged in Indicator 4.1.1.

Liberia's experiences seem to confirm the need for clearer guidance on how to deal with examinations, relative to assessments, as a potential source of monitoring data. In 2007, Liberia was planning to introduce a sample-based national assessment, but better use of data from the Grade 6 national examinations to gauge progress was also envisaged as an interim measure[46]. The examination was subsequently dropped, but in 2016 the government was planning to re-introduce the examination (the 2016 plan still put forward plans for a sample-based national assessment).[47]

Clearly, one advantage with the use of examinations for gauging trends is that they already feature prominently in the policy debates of many countries. Stakeholders are familiar with them, and understand them, even if they often over-estimate the comparability over time of the statistics. Examinations undoubtedly provide some guidance to policymakers and the public in relation to the extent to which children are not acquiring basic skills. They are certainly better than having nothing.

If examinations data are to be used at all, even as interim data sources, for reporting Indicator 4.1.1, it is likely that their data have to be used in new ways. Examination results often reflect aggregate outcomes across several school subjects. There would need to be a stronger emphasis on extracting statistics on the two fundamental learning areas of Indicator 4.1.1. Trends over time have to be interpreted with considerable caution. Adjustments to raw results may be necessary to improve the comparability of the statistics. In South Africa, such adjustments, which made use of a group of presumably stable schools as an anchor, changed a downward trend to an upward trend.[48] More demanding examinations seemed to lie behind the original downward trend. However, these types of adjustments are technically complex, defensible only to a limited

[44] South Africa: Department of Basic Education, 2016, p. 33.
[45] Sri Lanka: Department of Examinations, 2015.
[46] Liberia: Ministry of Education, 2007, p. 23.
[47] Liberia: Ministry of Education, 2016.
[48] Gustafsson, 2016.

degree, and may not be widely believed by the public. Costs in this regard would have to evaluated against the costs of establishing a new sample-based national assessment system, which could provide statistics without these complexities.

In the light of the above discussion, a five-level hierarchy is proposed for this report. These five levels are used in Section 3, where the geographical distribution of different assessments systems is described. The five levels are, from more to less reliable: (a) the three large international programmes (PISA, TIMSS and PIRLS); (b) five regional cross-national assessments; (c) sample-based national assessments; (d) censal assessments; and (e) national examinations.

The distinction between (a) and (b) is fairly standard, as is the distinction between (b) and (c). What is novel about this hierarchy is that it identifies sample-based national assessments as preferable to censal national assessments – (c) versus (d). Sample-based national assessments seem more likely to follow a standard recipe, in many ways one established by the cross-national programmes. Censal assessments, on the other hand, come with security risks, and without widely accepted global design standards. They may provide reliable trend data, but it is also very possible that they do not. The hierarchy moreover includes examinations as a possibility should an assessment programme not be available. As suggested previously, the preferability of (d) over (e) may not always apply. A well-run examination whose statistics are carefully used to produce subject-specific proficiency statistics may be better than a poorly designed censal programme that lacks secure anchor items.

This section does not get close to providing a complete account of the many factors that should be taken into account when decisions are made about what assessments to use when reporting on progress in basic proficiency. The points discussed above could inform the further development of the Principles of Good Practice in Learning Assessment (GP-LA),[49] a guide developed for the UIS. That guide, though providing a useful framework, should ideally include more guidance on matters such as the trade-offs between sample-based and censal national assessments, the costs and benefits of having test designs that follow a matrix approach, optimal test administration processes, and when (if ever) examinations provide an opportunity for gauging trends.

2.7 The extent and clarity of the out-of-school phenomenon

Key points: Though Indicator 4.1.1 is formally only concerned with the proficiency levels of children in schools, it is crucial that the proportion of out-of-school children be taken into account when interpreting Indicator 4.1.1 values.

It is important to note that Indicator 4.1.1 is meant to describe children and youths who are attending school, not those who for some reason are not attending.[50]

A methodology has been developed for the UIS for a global composite indicator which combines the effect of non-proficiency among those attending school with those not attending school. It is estimated that in

[49] UIS, 2017j.
[50] UIS, 2017e, p. 8.

recent years around 60% of children of primary school age are not achieving desired the proficiency levels in reading, with the figure being similar for mathematics. Of these non-proficient children, around two-thirds reach the final grade of primary school, but do not learn enough to attain the desired proficiency benchmarks.[51] Thus, overall 40% of children become proficient, 40% complete primary school but as non-proficient students, and 20% of children do not complete primary school (and are assumed not to reach the proficiency benchmark associated with the end of primary).

It should be noted that this type of global accounting comes with several uncertainties which should be made clear and dealt with as far as possible. For instance, for many countries the percentage of children completing primary schooling varies considerably depending on the data source and method of calculation. Moreover, UIS and UNICEF statistics on net enrolment ratios have differed considerably in the past.[52] Especially in developing countries, the number of repeaters in the grade being tested is often high. Depending on how this is dealt with, and depending on the performance of repeaters relative to non-repeaters, grade repetition could result in an over-estimate or an under-estimate of actual proficiency levels. Of course, this estimation problem is not really affected by the number of out-of-school children, and the problem actually affects the comparability of statistics discussed in Section 2.4. However, this problem may best be dealt with at the point when global composite indicators are calculated.

2.8 The timeliness, credibility and policy impact of the statistics

Key points: Assessments produce national, and often sub-national, statistics which can influence policymaking and policy implementation in positive ways. For these positive impacts to be felt, statistics must not only be accurate, they must be widely *seen* to be credible, and the turnaround time between the assessment and the reporting of results should be as short as possible, without compromising on quality.

In previous sections the emphasis has been placed on having reliable statistics, and understanding what these statistics tell us, and how (if at all) one can compare statistics over space and time. In the area of learning outcomes, however, it is not enough for the experts to agree on the statistics. There needs to be public buy-in. The understanding of the experts, and sometimes their reservations, needs to be communicated widely.

The fact that assessment programmes and their results can be controversial, far more so than other education matters such as enrolment or financing, makes good communication strategies vital. Education International, the world federation of teacher unions, and some education academics, have been highly critical of cross-national and national assessments.[53] In part, this criticism stems from concerns with core features of typical assessments, for instance their focus on a relatively narrow set of competencies, but in part the criticism relates to concerns around the reliability of results. Criticism tends to be stronger when assessments are censal, as opposed to sample-based, as censal assessments are virtually all used to hold schools accountable in some way. Poorly designed accountability systems can lead to allegations that these systems unfairly find fault with individual schools. Moreover, assessments are often opposed on ideological

[51] UIS, 2017a; UIS, 2017c, pp. 5, 11.
[52] Gustafsson, 2015.
[53] Education International, 2011.

grounds. They are seen to respond purely to the needs of business and the private sector, for instance. It is vital that the value of assessments as public goods, and as tools that can highlight and reduce inequalities, be continuously emphasised in, for instance, the reports of UNESCO.[54]

Communication strategies should explain why learning assessments are necessary, but should also clarify that different assessments serve different purposes, and have different challenges. This is necessary to prevent an across-the-board rejection of assessments. A problematic assessment programme in one country should not make programmes in all countries problematic.

It needs to be explained that trend data on learning outcomes can help countries to understand whether their education policies are working in an overall sense. The background questionnaires often attached to assessments, in particular sample-based ones, permit some insights into what aspects of the schooling system could be holding back progress, though these data almost never provide conclusive evidence of what is or is not working. Assessment programmes also help to identify inequalities, for instance those between boys and girls, or between rural and urban areas.

The importance of transparency and public accountability in assessment programmes is often emphasised. A 2017 report for the UIS makes this point as follows:[55]

> All aspects of an assessment programme should be open to outside scrutiny. This means that the assessment methodology, implementation processes and data analysis methods and procedures should be clearly described and publicly available. By justifying the decisions made in relation to the assessment methodology, implementation and analysis, the results are not only verifiable by other experts in the field, but they are more robust to criticism. This also helps contribute to the objectivity of the results.

Making the technical aspects of a programme transparent thus promotes technical rigour, and helps to counter the impression that the assessment is creating a distorted picture for political purposes. Many countries make the microdata from their national assessment programmes publicly available, or available to external analysts with credible research proposals. This can promote public trust in the assessment programme. However, this usually requires technical skills within the assessment authority in order to anonymise data, meaning the removal of information which could link data to individuals or (in the case of sample-based assessments) individual schools.

The temptation to hide flaws in the programme, in the case of a national assessment system, can be significant. Exposure of these flaws could be embarrassing for an assessment authority, and might even put the future of the programme at risk. The appropriate advice is probably the following. Yes, there are risks associated with revealing flaws, but the arguments in favour of transparency remain strong. A lack of transparency can result in a situation where critical flaws are not known, or at least not understood, even by experts within the assessment authority. This increases the possibility that the programme will produce misleading statistics for many years into the future. A compromise can be to allow maximum transparency

[54] Benveniste (2002) provides a rare example of an academic input advancing the notion of assessments as a public good. Specifically, he argues that effective assessments are an integral part of the welfare state.
[55] UIS, 2017j, p. 7.

for a limited group of experts, from inside and outside the assessment authority, to ensure that criticism of the programme remains technical and constructive, as opposed to politically driven.

The timeliness of the results from the assessment programmes is important. Currently the lag between testing and the publication of results is about a year for PISA, TIMSS and PIRLS. The lag varies considerably for the regional programmes: one year for the PASEC 2014 results to be published, three years for the LLECE 2013 results, and four years for initial reports for specific countries participating in SACMEQ 2013 to emerge. It is not clear what the variation in the lags of national assessment programmes is. Here it is likely that censal programmes will have a faster turnaround time. Despite being larger than sample-based assessments, pressure to have final results per school available soon would be strong. Sample-based assessments, on the other hand, experience less direct time pressures. This has advantages – there is more time for rigorous verification of the data – but also disadvantages. Policy should be informed by data which is as recent as possible. Better information on the turnaround times of national assessments, which could be collected through the UIS questionnaires, could help countries decide whether their national assessments were unreasonably slow in producing final data. Moreover, organizations such as the UIS could help to improve the utility of data from the regional programmes by advising on how these programmes could all attain the one-year turnaround time of, for instance, PISA and TIMSS.

2.9 Incentives within the UIS collection system

Key points: The UIS aims to help countries improve their national data collection for Indicator 4.1.1 , by providing them with services such as capacity building tools, feedback relating to the reliability of data collection systems and policy advice to facilitate cross-country comparison of statistics. However, criticism of national or even cross-national assessment systems could undermine collaboration with the UIS. Constructive criticism is vital, but how it is packaged can have large implications.

Countries participate in the UIS reporting systems on a voluntary basis. Unlike a typical national government collecting statistics from, for instance, provinces, the UIS has no legal mechanism to deal with non-compliance or poor quality data. It is thus not surprising that there are serious gaps in the UIS datasets. For example, in 2018 only around 55% of countries had some value for the years 2013 to 2015 on total spending on education in UIS.Stat. The corresponding figure for primary enrolment was 80%, so even for this basic statistic there are serious gaps. A 2017 audit of the availability of SDG statistics found that only around half of all countries were able to report all 11 global indicators for SDG 4.[56]

Willingness to participate in the UIS reporting system needs to be understood in terms of soft incentives. This could include being seen to be cooperating internationally. Countries must also see the benefits and usefulness of the statistical products emerging from the UIS using the submitted data. These incentives are particularly complex in the case of proficiency statistics, given the politics and measurement difficulties they present.

One can think of four categories of tools produced by the UIS, which assist countries in various ways, and serve as incentives for countries to participate. Here the focus is not just on proficiency statistics, but on the wider system of UIS education statistics.

[56] UIS, 2017k, p. 19.

Firstly, materials which can be used to build capacity, and interactive training by the UIS using these materials, are important (the importance of building human capacity was discussed in Section 2.2).

Secondly, guides and manuals can assist countries in gathering and processing data. Here the suite of country questionnaires used by the UIS to gather information on an annual basis is crucial – these tools are arguably used more widely than any other UIS tool.[57] The guides that come with these questionnaires provide basic guidance, but crucially do not help countries deal with imperfect data, which arguably is a problem in most countries. Moreover, the questionnaires are limited when it comes to gathering metadata, specifically inputs by countries relating to possible reliability problems. The system is also not set up to accommodate corrections in earlier years. Allowing for these things would clearly make the questionnaires more complex, but conversations with users suggest that enhancements would be worth this cost. The challenges mentioned here will be magnified when and if the questionnaire system is expanded to collect data from national assessments. The importance of the design of these tools goes beyond the international reporting system. The way collection systems within countries work appears to be strongly influenced by international systems. This means that problems in the UIS system – such as a lack of guidance on how to work with imperfect data – are easily replicated within national systems. But it also means that innovations in the UIS system are likely to have a positive impact on within-country systems.

Thirdly, a UIS review of national data collection systems is a valuable tool which can be used to improve not just the reviewed countries, but countries in general.[58] This incentive in the UIS system is optimised if a review goes beyond simply looking at compliance and processes, and examines the quality of data, in part through analysis of the degree of consistency across statistics, and in part through checking microdata. Obviously this level of review carries significant costs, meaning selecting specific countries as case studies becomes important. Apart from financial costs, there are political risks. A country may agree to have its microdata scrutinised, but then oppose the review if the UIS finds serious problems in the data – which could embarrass a government. In such instances, the UIS could provide a confidential report to the government, and make public a separate report more oriented towards general constructive criticism. For these types of case studies to be cost-effective, some level of public reporting is necessary.

Fourthly, monitoring and policy-focussed reports providing international comparisons are a vital tool for policymakers in countries. An example of this would be a 2017 UIS report on regional disparities with respect to proficiency.[59]

In all the tools discussed above, displaying a clear awareness of the limitations of the statistics and the underlying data is vital. This is especially true in the case of proficiency statistics. Without this awareness, the risk increases that countries will withdraw from the new systems that report on proficiency. Two key risks can be identified. One relates to comparisons across countries, the other to comparisons over time. In a scenario where statistics are derived from national assessments, a country may feel that its ranking is unreasonably low because it applies more stringent and ambitious proficiency benchmarks. In short, a country could feel it was being punished for being ambitious or honest, and could reject the international reporting system. In relation to comparisons over time, a country's trend may seem dubious, perhaps because the national assessment is still weak, and as a result a country may wish to stop submitting statistics.

[57] http://uis.unesco.org/en/methodology, under heading 'UIS Questionnaires' (accessed June 2018).
[58] UIS, 2016.
[59] UIS, 2017c.

This is more likely to occur if there are not clear warnings in the UIS systems and reports to prevent inappropriate interpretations.

A specific problem that is likely to arise from the use of national assessments is improvements over time which, in the view specialists, are unbelievably large. The political incentives to believe such trends, and not to take steps to investigate the data, could be strong. Here the UIS could assist by developing guides on what degree of improvement can be considered a 'speed limit', in the sense that improvements beyond this degree would be unlikely, given historical trends. There is now a sizeable literature on how large an improvement one can expect, in terms of standard deviations, from an exceptionally good intervention applied to a sample of schools. However, historical trends for whole countries suggest that countries are unlikely to see such large improvements. What exceptional countries can reasonably expect, based on trends seen among the strongest improvers of the past, could be made clearer. Hanushek and Woessman (2007, p. 44) provide some guidance, but this could be updated using more recent data, and should ideally differentiate to a larger extent between countries at different levels of development. This type of guidance, apart from assisting in the interpretation of national trends, could help countries set targets, and help UNESCO gauge how easily the SDG goal of quality schooling for all children could be achieved.

3 Understanding the current configuration of assessments

Knowing what currently exists in countries with respect to assessment systems is important for charting a way forward for Indicator 4.1.1. There have been a few audits, but these are arguably incomplete for the purposes of assessing the costs and benefits of the proposals outlined in Section 4. For instance, a recent UIS investigation[60] into the costs of reporting against SDG 4 illustrates the extent to which countries participate in cross-national assessments, or have national assessments, but this is not broken down by education level or learning area. If coverage of an education level that is considered strategically important is particularly low for specific groups of countries, this could influence UIS strategies. The UIS online database (UIS.Stat) indicates which countries have a 'nationally representative learning assessment', with data broken down by the three education levels and two learning areas. However, this database does not indicate whether learning assessments are cross-national or national, and (in the latter case) sample-based or censal, issues which could have implications for the comparability of statistics over time *(see Section 2.6).*

Fortunately, there are various information sources one can use in order to piece together a picture of potential Indicator 4.1.1 country-level data sources, in a manner that assists the discussions of the current report. In particular, the UIS 'Database of learning assessments' (which is separate from the main UIS.Stat system) provides useful information about the national systems (including examinations) of certain developing countries. A 2013 OECD report[61] offers a useful account of systems in developed countries. Information on cross-national programmes, at least as far as their coverage is concerned, is easily available in various reports.

The current section uses the various available sources to describe two configurations of assessments in a manner that assists the analysis. One represents systems that have produced some kind of statistics on learning outcomes over the last five or so years. This is called the 'already realised coverage'. Importantly, only nationally representative statistics were considered, meaning that statistics representing regions in

[60] UIS, 2017h, pp. 17, 33.
[61] OECD, 2013, Annex 4.A2.

countries were not counted. A second configuration, called the 'optimistic near future coverage', takes what exists in the 'already realised coverage', and adds programmes, or the participation of new countries in existing cross-national programmes, where the future generation of information is more or less assured, or where recent trends or political commitments suggest that information might be generated in the near future (roughly within the next five or so years).

The focus falls strongly on differentiating coverage in terms of, firstly, the three education levels and, secondly, the five types of assessments outlined in Section 2.6. Less attention goes to whether both learning areas are covered. The reason for this is that one can assume that adding a learning area in those rare instances where only one is covered *in a national assessment* would be relatively easy to achieve. What is thus more critical is whether national testing exists at all in a specific country and at one of the education levels. Filling gaps in this regard is more challenging, and one may have to plan for an extended existence of these gaps. With regard to the cross-national assessments, both reading and mathematics are covered by virtually all systems, with one notable exception: TIMSS covers only mathematics and PIRLS covers only reading. The learning area gaps arising as a result of the TIMSS-PIRLS situation are discussed below.

Table 1 and Table 2 report on the already realised coverage. The first table considers coverage in terms of the total populations[62] of countries, while the second table simply counts countries.[63]

For TIMSS and PIRLS, participation in either of the two most recent waves was counted (2011 and 2015 for TIMSS, 2012 and 2016 for PIRLS). In the case of PISA, participation in 2012 or 2015 was considered. What are the learning areas and education levels covered by these assessment programmes? PISA is clearly mostly lower secondary, though one can assume that in many developing countries large proportions of the PISA age 15 target group would still be at the primary level. PISA covers both reading and mathematics (and also science). TIMSS covers mathematics (and science) in Grades 4 and 8. Of participating countries in TIMSS, 60% have a primary school cycle consisting of six grades, and only two countries have an eighth grade at the primary level, so Grade 8 can be considered being virtually always at the lower secondary level, in fact mostly the second year of secondary schooling. Thus TIMSS Grade 8 has been attached to the lower secondary level. Grade 4 is the end of the primary cycle for just 9% of TIMSS countries, so arguably TIMSS Grade 4 would be useful in the SDG sense mainly as an indication of learning that occurred at the lower primary level, so the (a) of Indicator 4.1.1. The same can be said of PIRLS, which with a few exceptions focusses on Grade 4, the difference being that PIRLS assesses reading.

To a large degree, both learning areas would be covered by the IEA's two programmes (TIMSS and PIRLS): only 4% of the global population is in countries with TIMSS in Grade 8, but no PISA, which assesses reading at roughly this level (this is largely accounted for by Egypt, Iran and South Africa). At Grade 4, TIMSS covers 28% of the world's countries and also 28% of its population. At the Grade 4 level, 5% of the world's population (largely Japan, Republic of Korea and Turkey) has TIMSS but no PIRLS, meaning mathematics but no reading coverage, while for 2% the reverse of PIRLS and no TIMSS applies. The three large international assessments have no footprint at the end of primary level.

[62] Population totals are the most recent ones per country on the UIS online database.
[63] 223 countries were assumed to be the maximum and includes the 224 'countries' in the UIS online database tables on SDG indicators, minus 'Sudan (pre-secession)'. Sudan and South Sudan as separate countries were thus counted.

Table 1. Assessment type and world population coverage (percentages)

	Lower primary	End of primary	Any primary	Lower secondary	Any level
PISA, TIMSS and PIRLS	28	0	28	38	38
Regional cross-national assessments	12	16	16	0	16
Any cross-national	38	16	42	38	47
Sample-based national assessments	62	35	66	55	66
Censal assessments	9	11	12	7	12
National examinations	1	12	12	13	16
Any of the five types	89	50	94	91	96

Table 2. Assessment type and coverage by number of countries

	Lower primary	End of primary	Any primary	Lower secondary	Any level
PISA, TIMSS and PIRLS	62	0	62	84	87
Regional cross-national assessments	47	62	62	0	62
Any cross-national	106	62	119	84	137
Sample-based national assessments	47	41	62	25	62
Censal assessments	22	25	30	11	33
National examinations	4	29	30	35	41
Any of the five types	135	110	155	129	167

For the regional cross-national assessments, four programmes were counted: SACMEQ, LLECE, PASEC and PILNA.[64] Details on these four progammes, plus a fifth programme, SEA-PLN[65], are given in **Table 3.**[66] SEA-PLN is only counted for the 'optimistic near future' picture in a later table as this programme has yet to produce country statistics. All the five programmes cover (at least) the learning areas reading and mathematics. Their education levels, in terms of actual grades tested and assumed correspondence to Indicator 4.1.1 are indicated below. The five programmes appearing in Table 3 are also the five regional programmes considered by the IEA as candidates for the Rosetta Stone linking method described in Section 4.3.

For the line 'sample-based national assessments' in Table 1 and Table 2, a variety of sources were used. In line with the discussion above, it was assumed that for gauging national improvements over time, a sample-based assessment was a better measurement tool than a censal assessment *(see Section 2.6).* Above all, two publicly available sources were used: the UIS Database of Learning Assessments,[67] covering 68 developing countries, and Table 4.A2.5a of OECD (2013), which reflects standardised central assessments at the primary

[64] Pacific Island Literacy and Numeracy Assessment.
[65] Southeast Asia Primary Learning Metrics.
[66] The past 14 SACMEQ countries are listed in Makuwa (2010). The planned participation of Angola is indicated at http://www.sacmeq.org. The 17 LLECE countries are countries listed either in UNESCO (2008) or Flotts et al (2014). Fifteen PASEC countries are listed at http://www.pasec.confemen.org/evaluation/evaluation-internationale-pasec2019. A 16th country, Comoros, participated in 2009 according to CONFEMEN (2012), and has been counted here. The 11 SEA-PLN countries are listed at http://www.seaplm.org. The 14 PILNA countries are listed in Belisle et al (2016).
[67] http://uis.unesco.org/en/uis-learning-outcomes, accessed June 2018.

level in OECD countries. The decision as to whether to classify a national assessment as being at the lower primary level or the end of primary was influenced in part by separate UIS data on the duration of the primary cycle. It was also influenced by the number of grades assessed within a country. The education level classifications used are in part debatable. However, one should keep in mind that the intention here was to produce a general picture of the coverage, not to make firm proposals on which grade-specific assessment to consider for Indicator 4.1.1.

Table 3. Details on five regional programmes

	Region and countries counted for the graphs	First system-wide assessments	Educational levels
SACMEQ	Southern and East Africa (15)	1991	Grade 6, so the end of primary education.
LLECE	Latin America and Caribbean (17)	1997	Grades 3 and 6, so lower primary and end of primary education.
PASEC	Francophone Africa (16)	1993	Grades 2 and 5, which can be considered to correspond to the (a) and (b) of Indicator 4.1.1 (lower primary and end of primary education).
SEA-PLN	Southeast Asia (11)	After 2017	Grade 5, which for three countries is the end of primary education (Grade 6 for the others). The education level is thus considered end of primary education.
PILNA	Pacific Islands (14)	2012	Grades 4 and 6, which has been assumed to correspond to levels (a) and (b) of Indicator 4.1.1.

A more serious problem than classifications is missing data. There are clearly several countries which have been excluded from the two information sources, presumably in part because the relevant information could not be found. Moreover, the two sources do not reflect new assessments introduced in the last few years. A few obvious gaps for large countries were filled. Sources indicate that China ran its first sample-based national assessment in 2017, with results available for two targeted grades, Grades 4 and 8 (which are mapped onto lower primary and lower secondary in this analysis, Grade 6 being the end of primary in China).[68] Also in 2017, India conducted its National Achievement Survey for the first time, testing samples of Grades 3, 5 and 8 students (considered here representative of the three Indicator 4.1.1 levels), and publishing national and state-level results.[69] Brazil's national testing system, started in 1990, now collects data from Grades 3, 5 and 9, on both a sample-based and census basis.[70] Brazil can thus be considered to cover all three Indicator 4.1.1 levels with its national assessments (which were entered as just sample-based systems for the current analysis, meaning the censal component was ignored). In the United States, the NAEP

[68] China: Ministry of Education (2017), also UNESCO (2017, p. 127).
[69] http://mhrd.gov.in/NAS, under heading 'National Achievement Survey - 2017'.
[70] https://www.somospar.com.br/saeb (in Portuguese).

programme assesses samples of students in Grades 4 and 8 (so lower primary and lower secondary).[71] Of course, in the case of Brazil and the United States, the national systems do not affect coverage insofar as what their national systems cover are also covered by cross-national systems. A 2013 report indicates that at least at that point in time, there was no national assessment system in Russia.[72]

For censal assessments and national examinations, the two lists of assessments compiled by UIS and OECD mentioned previously were used (though only the UIS list includes national examinations). These two categories of assessment would also be incompletely covered insofar as both lists are clearly missing information for some countries.

UIS.Stat had (in 2018) values indicating whether a country had a nationally representative learning assessment in 2015, or in the previous five years,[73] with indicator values broken down by the three Indicator 4.1.1 education levels and two learning areas. These values are supposed to count both cross-national and national programmes. The UIS figures correspond roughly with the figures seen in the analysis presented here, and to a degree this confirms that the picture provided above can inform the strategic choices that must be made. Specifically, the UIS values are a bit lower than the values provided in Table 1. For instance, where the bottom line of Table 1 indicates that 89%, 50% and 91% of the three education levels are covered in population-weighted terms, the corresponding percentages using the UIS values would be 56%, 56% and 79%. The UIS figures thus also point to the best coverage being at the lower secondary level. The fact that the lower primary value is so much higher in Table 1 compared to what one obtains using UIS values is largely due to the fact that China's emerging national assessment at this level has been counted here (but not in the UIS database).

What are the key patterns emerging from Table 1 and Table 2? The first table indicates that 96% of the population-weighted world has some type of assessment at some education level. In reality, this figure is likely to be 100% as there would be few schooling systems with no national assessment or examination at all (there are some, for instance Angola, at least up till recently). Much of the missing information problem relates to small countries: the 56 countries (the total of 223 countries minus 167 from Table 2) representing 4% of the world's population. It is encouraging that 94% of the world's population has some kind of assessment at the critical primary level from which the global reporting system could potentially draw (see Table 1). Moreover, almost one-half of the world (47%) has been participating in some cross-national programme (this figure becomes higher if one considers cases where parts of countries participate in these programmes). This does imply that for any one of the three education levels, over half of the world would have to be monitored through the use of national assessments (or examinations), at least given current levels of coverage of the cross-national programmes.

Table 4 provides critical information not provided by Table 1 and Table 2 (though the bottom row of Table ***4*** is the same as the bottom row of Table 1). Here the best possible assessment type, using the hierarchy discussed previously, is considered. We see, for instance, that 38% of the world's population is covered by some type of cross-national assessment at the lower primary level, but also the lower secondary level. The figure is a much lower 16% for the end of primary. Including sample-based national assessments makes a very large difference to coverage. At the lower primary level, for instance, the coverage rises from 38% to

[71] United States: Department of Education, 2013.
[72] Tyumeneva, 2013.
[73] UIS, 2017e, p. 9.

89%. This is largely due to China's and India's national assessments. Including censal national assessments does not make a very large difference to global coverage, and virtually all of this difference is at the end of primary level. Examinations increase the coverage, in particular at the lower secondary level. However, the main boost to global coverage is brought about by sample-based national assessment. This is encouraging, and the importance of distinguishing between sample-based and censal assessments becomes clear. Sample-based assessments are easier to manage and to improve when it comes to their capacity to gauge improvement over time.

Table 4. Summary for already realised coverage

	Lower primary	End of primary	Any primary	Lower secondary	Any level
PISA, TIMSS and PIRLS	28	0	28	38	38
+ regional cross-national assessments	38	16	42	38	47
+ sample-based national assessments	89	43	92	84	95
+ censal assessments	89	48	92	85	95
+ national examinations	89	50	94	91	96

Note: Values refer to percentage of the world population.

The relatively low coverage for the end of primary is important. This suggests that to some extent the focus needs to fall on drawing from any primary-level assessment, in other words to view primary as one category. In terms of understanding global policy challenges in broad terms, whether one uses lower primary or end of primary statistics does not make a huge difference. Note that the difference between 94% and 89% in the final row of Table 4 indicates that one will not maximise coverage at the primary level by focusing only on lower primary. There are countries with end of primary but not lower primary statistics.

Figure 6 illustrates the coverage by world region.[74] A pattern that stand out is the low coverage, in terms of countries, when it comes to lower secondary in the Pacific region. Much of this is likely to be the result of data gaps in the international databases with regard to national assessments and examinations in this region, as opposed to an actual absence of lower secondary assessments.

Figures 7, 8 and **9** illustrate the details behind Table 4. The approach in compiling the maps was to take the best possible assessment type per country. So what do these maps reveal? Figure 7 confirms an important disparity in Africa: Francophone countries covered by PASEC have statistics at the lower primary level, but this is not the case for the (mostly) anglophone SACMEQ countries. Hence for the SACMEQ countries it becomes necessary to rely rather heavily on national assessments at this level, and clearly many of these countries have sample-based programmes.

[74] The eight world regions are the main regions used in the statistical tables of UNESCO's 2017/18 *Global Education Monitoring Report*.

Figure 6. Already realised coverage by world region

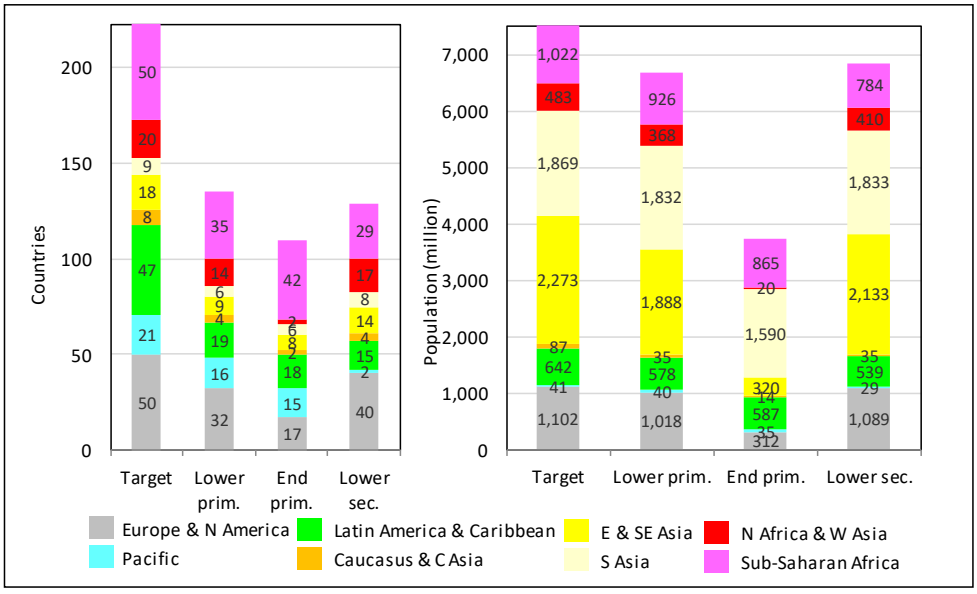


Figure 7. Already realised coverage of lower primary education

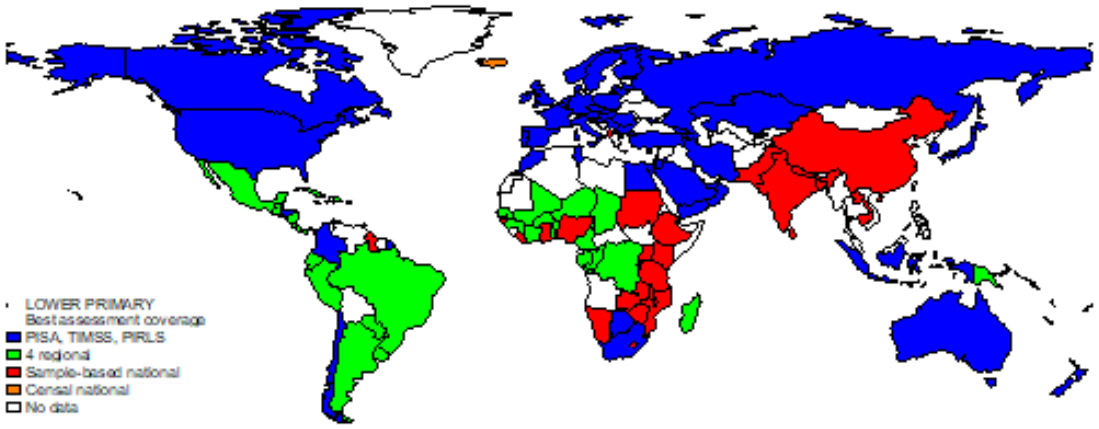


Figure 8. Already realised coverage of end of primary education

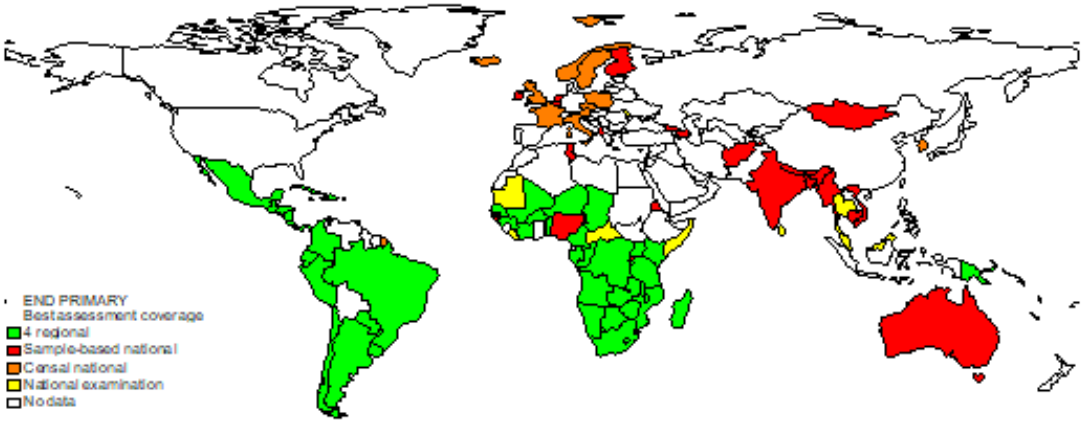


Figure 9. Already realised coverage of lower secondary education

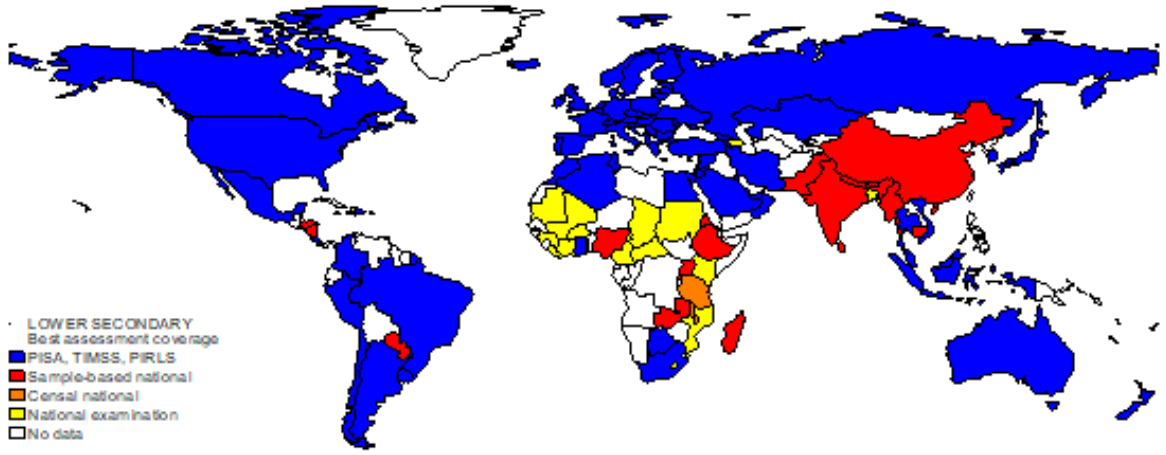


Figure 9 illustrates where national examinations do emerge as important: at the lower secondary level in many African countries. For these countries (in yellow) examinations data is the best there is, given the absence of assessments.

Attention now turns to the 'optimistic near future coverage'. **Table 5** illustrates this future scenario – all assessments counted for Table 4 are included, plus some additional participation in cross-national programmes. Three types of additions were made. Firstly, PISA 2018 participation was considered.[75] A major addition implied by this is the inclusion of China as a whole country.[76] In 2015, two provinces and two municipalities in China, representing 17% of China's population, participated in PISA. Moreover, a few other countries join PISA for the first time in 2018. What is not assumed to occur in the near future, is the

[75] http://www.oecd.org/pisa/aboutpisa/pisa-2018-participants.htm (accessed June 2018).

[76] China is listed as a 2018 participant on the PISA website, but what seems to confirm that China *as a whole country* is participating in 2018 is an explicit mention of this in the Wikipedia page for "Programme for International Student Assessment" (accessed July 2018).

participation of India as a whole in PISA, despite the fact that states within India have participated in recent years. Specifically, two states, representing 6% of India's population, participated in PISA in 2009.

Secondly, it was assumed that the participation of eight developing countries in the special PISA for Development programme would result in statistics for these countries at the lower secondary level.[77] Thirdly, it was assumed that SEA-PLN would proceed as planned.

The key coverage expansion brought about by the optimistic scenario is at the lower secondary level and at the end of primary level – the latter rises from 50% (bottom row Table 4) to 55% *(see Table 5)* due to the changes in Southeast Asia brought about by SEA-PLN. Moreover, there is a shift from national assessments to cross-national assessments, as the two cross-national programmes PISA and SEA-PLN become the new best possible in several countries. For instance, at the end of primary level cross-national assessments rise from 16% of the population-weighted world to 25%.

Table 5. Summary for optimistic near future coverage

	Lower primary	End of primary	Any primary	Lower secondary	Any level
PISA, TIMSS and PIRLS	28	0	28	59	60
+ regional cross-national assessments	38	25	47	59	68
+ sample-based national assessments	89	50	95	86	97
+ censal assessments	89	54	95	88	97
+ national examinations	89	55	95	93	97

Figures 10 and **11** illustrate the end of primary and lower secondary levels, the two levels which change following the additions.

Figure 10. End of primary optimistic near future coverage

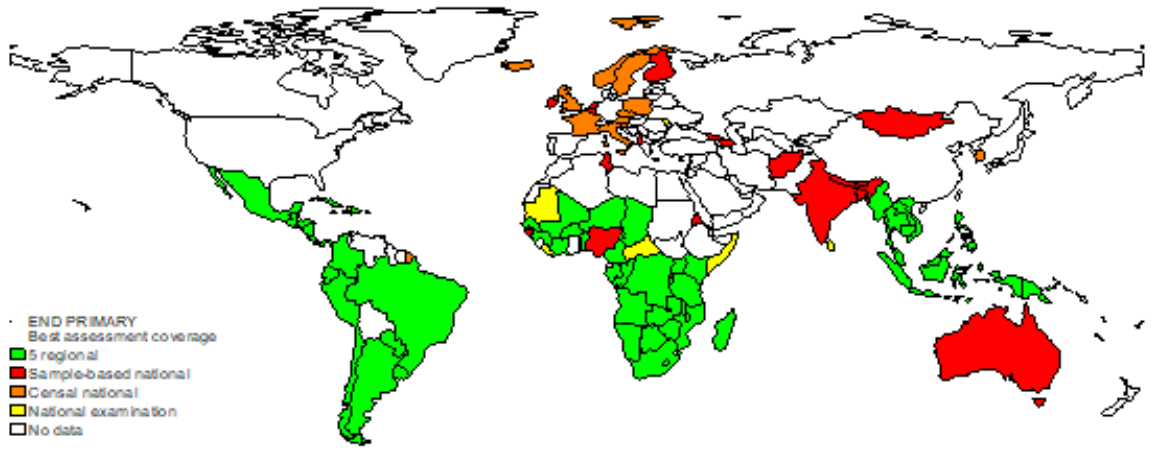
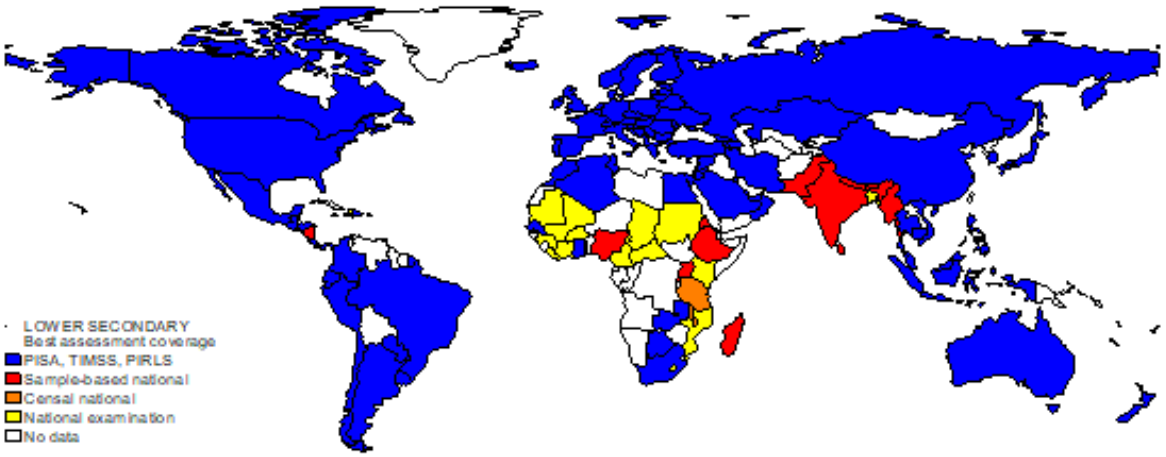


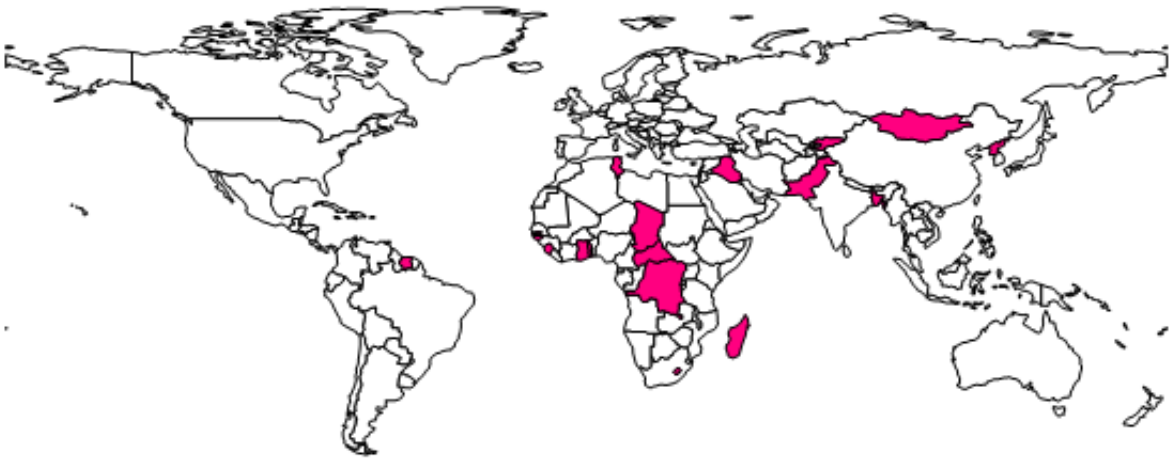
Figure 11. Lower secondary optimistic near future coverage

[77] http://www.oecd.org/pisa/aboutpisa/pisa-for-development-participating-countries.htm, under "PISA for Development - Participating Countries" (accessed July 2018).



Finally, it should be pointed out that international programmes that collect learning outcomes data from children through household surveys can serve as a vital reality check when data derived from schools-based programmes are evaluated. Household-based data is generally not prioritised as a primary source for reporting Indicator 4.1.1. This seems partly due to the fact that the position of UNESCO, and other leaders in the education sphere, is that schooling systems should all have assessment systems as an integral part of their improvement mechanisms. One disadvantage with permitting the use of household-based data for Indicator 4.1.1 would thus be an undesirable shift away from the core focus of establishing effective assessment programmes within a schooling systems. Yet household data, where available, ought to be used when schools-based data are validated. A major development has been the inclusion of reading and mathematics tests in Version 6 of UNICEF's Multiple Indicator Cluster Surveys (MICS).[78] The **Figure 12** marks the 18 countries implementing MICS with the learning assessment component in the 2017 to 2018 period[79]. This assessment component is designed for household members aged 5 to 17.

Figure 12. 2017-2018 MICS participants using learning assessments



[78] http://mics.unicef.org.

[79] The list of 18 countries was obtained from UIS.

4 Existing proposals for the way forward

What has the UIS committed itself to? Despite the availability of various analyses of the options, many published by the UIS, at present the UIS still maintains a fairly open position with regard to the way forward for Indicator 4.1.1. In a 2017 annual report, it is acknowledged that "it will take considerable time and resources to resolve the methodological issues related to Indicator 4.1.1".[80] The report justifies an incremental strategy as follows:[81]

> *…the most feasible approach in the medium term lies in linking cross-national assessments to report the results while continuing to develop more sophisticated tools, such as the UIS Reporting Scales. … At the same time, there is a wider challenge that goes beyond the technical and consensus-building work of GAML [Global Alliance for Monitoring Learning]. In particular, many countries do not want to participate in cross-national assessments and, for different reasons, are not conducting their own national assessments. These countries should not be left behind in the SDG reporting process. It is, therefore, essential to strengthen support by donors for learning assessment. The UIS is helping to highlight these issues by building an investment case for learning assessment.*

Here both cross-national assessments and national assessments are viewed as parts of the overall solution.

A key requirement for the current report is that it evaluate the costs and benefits, mostly in a qualitative sense, of three proposals for the way forward. Their titles for the purposes of the current report are given in the following box. They correspond to a set of four strategies presented as part of a 2017 overview of cross-national programmes (by Ernesto Treviño and Miguel Órdenes) and the strategies are indicated in square brackets.[82] Details pertaining to these four strategies are discussed under the headings of the three proposals.

- Statistical recalibration of existing data [part of Strategy 1]
- Pedagogically informed determination of cut scores ('social moderation') [close to Strategy 2]
- Recalibration through the running of parallel tests (Rosetta Stone) [close to Strategies 3 and 4]

The sequence of the three proposals is not accidental. The sequence represents, roughly, a move from lower to higher financial costs (at least as far as UIS is concerned), and at the same time a move, on the whole, to greater reliability in the statistics.

4.1 Statistical recalibration of existing data

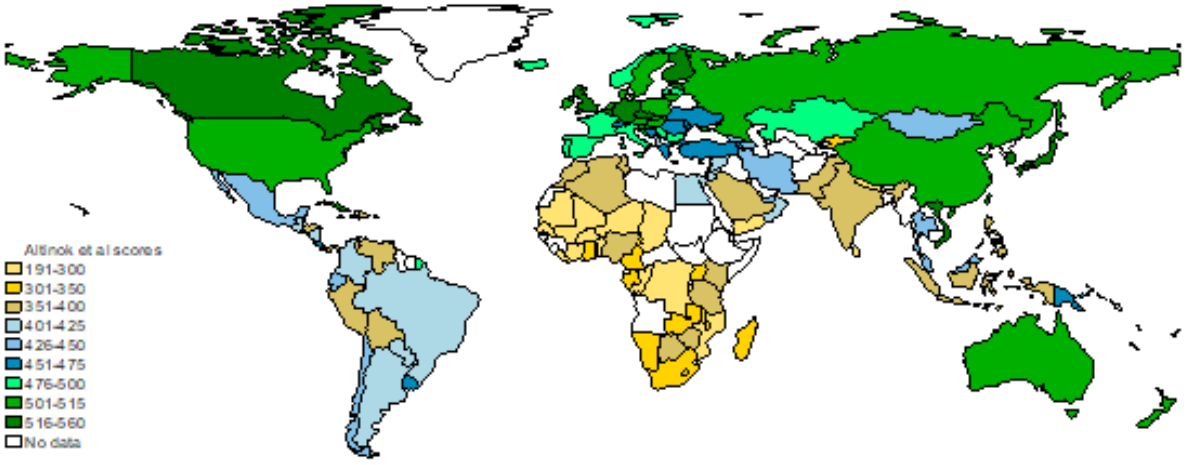
This proposal currently relies largely on statistical adjustments applied by Nadir Altinok to data emerging from cross-national programmes.[83] These adjustments take advantage of the fact that some countries, referred to as doubloon countries, participate in more than one cross-national programme. Using several

[80] UIS, 2017f, p. 12.
[81] UIS, 2017f, p. 13.
[82] UIS, 2017d, p. 24.
[83] Altinok, 2017; Altinok, Angrist and Patrinos, 2018.

such overlaps has allowed for the identification of roughly comparable cut scores representing global proficiency benchmarks across different programmes, as well as the calculation of confidence intervals around the resulting proficiency attainment statistics. The UIS.Stat tables that draw from Altinok's work provide values for 132 countries, the number of countries for the three education levels being 92, 48 and 97 (this is below the values 106, 62 and 84 of the middle row of Table 2 in part because assessment results from PILNA would not have been available). Note, however, that the values published on UIS.Stat are values before any statistical adjustments applied by Altinok.

Altinok's proficiency statistics per country appear not to be published on the web yet. However, Altinok, Angrist and Patrinos (2018, p. 77) includes the average scores per country, for the primary and secondary levels, and for a combination of the two, obtained through the adjustments process. Figure 13 below reflects the primary-secondary combination scores. The time periods per country would differ as all available data from the period 1965 to 2015 were used. Data from the Monitoring Learning Achievement (MLA) programme, a pre-2000 programme of UNESCO, were included in the set of data used. On the whole, the picture appears intuitively correct, though the rankings of certain countries would be considered surprising by many: China exceeding countries such as France; Papua New Guinea exceeding countries such as Thailand; Nigeria exceeding countries such as Ghana and Uganda. The case of China is easily explained by the fact that this country's PISA participation involved a relatively prosperous 17% of the country. Though the temptation may be great to use all the available data in estimating country scores, clearly there are data which are not sufficiently comparable for this purpose. It is better to exclude such data, or report separately on them, than to compromise on the credibility of the global dataset. Even with exclusions, it is likely that caveats in relation to specific countries will remain. Such caveats ought to be pointed out. This occurs only to a very limited degree in the two reports that inform this proposal.

Figure 13. Altinok et.al. recalibrated country scores



In calculating proficiency statistics, Altinok uses two proficiency benchmarks, or cut scores, one more lenient (and thus lower) than the other. This produces, after the statistical recalibration process, two alternative proficiency statistics per country, learning area and education level.[84] In making the separation between

[84] Altinok, 2017, pp. 47, 60, 61.

lower primary and end primary, so the (a) versus (b) separation in Indicator 4.1.1, Altinok considers Grades 2 to 4 as belonging to the former, and Grades 5 and 6 belonging to the latter.

Altinok is clear that his recalibration of existing data is a second-best approach, and that the ideal is the collection of data using standard data collection instruments.[85] It is clear that the proficiency statistics calculated by Altinok can often not be used as one would want to use such statistics, for instance to gauge a country's progress over time, due to the margin of error around the statistics. In a separate UIS report, the risks of this type of statistical recalibration approach are spelt out.[86]

> *… the complexity of the task of either trying a statistical linking or an equating among assessments may involve enormous economic, technical and transaction costs to obtain limited levels of validity in measuring Indicator 4.1.1.*

Yet the low costs of this statistical adjustments approach, at least compared to the other two proposals described below, could make this approach appealing. Moreover, at present this approach has produced what are arguably the most internationally comparable statistics in relation to Indicator 4.1.1. It is thus vital to make the costs and benefits of this approach clear.

Treviño and Órdenes, in their Strategy 1, see the utility of the statistical recalibration approach in its ability to provide a reality check against which to compare statistics based on national assessments.

4.2 Pedagogically informed determination of cut scores (social moderation)

What is discussed here as a single proposal is actually two separate, but overlapping, proposals put forward by the Australian Council for Educational Research (ACER, 2017) and Management Systems International (MSI). MSI has produced two reports for the UIS on the social moderation approach.[87] The term 'social moderation' is put forward by MSI, but not ACER. MSI also attaches the term 'policy linking' to their proposal.

What do the ACER and MSI approaches have in common? Both involve work by a team of experts to arrive at comparable cut scores in the various cross-national assessment programmes, at least in part through pedagogically informed evaluation of documents from the existing programmes. A key outcome of this work would be a tool which would describe the meaning of the cut scores, or proficiency benchmarks, in terms of what students can do. This tool would be used by national assessment authorities to establish cut scores in national assessments to distinguish proficient from non-proficient students.

The ACER proposal refers to the tool as a reporting scale. This proposal is the more costly of the two, in financial terms but also in terms of delays, mainly because of what is referred to as a validation phase, which involves running new assessments among samples of students in various countries.

[85] Altinok, 2017, p. 3.
[86] UIS, 2017d, p. 11
[87] UIS, 2017g; UIS, 2017l.

The reporting scale, a tool which UIS has already publicly expressed a commitment towards,[88] would indicate a range of competencies, from less to more advanced, in terms of descriptions such as find the value of a simple algebraic expression.[89] The same reporting scale would be used for the three Indicator 4.1.1 education levels, with the competencies expected for a higher level being found higher up the scale. There would be different scales for mathematics and reading. A first phase, drafting the reporting scales has been completed. This involved first formulating a conceptual growth framework informed by both the literature on how children learn, but also how the curricula of countries are structured. Thereafter, items from existing cross-national, and a few national assessments were examined by specialists, using established procedures to combine the judgements of the experts, so that the difficulty of items could be ranked and they could be understood in terms of the growth framework. In addition, experts examined pre-existing values describing the difficulty levels of the items analysed – these would be along the lines of the item locations calculated in Rasch analysis. It appears some new analysis of item-level data from existing assessment datasets was also undertaken.

[88] UIS, 2017f, p. 12.
[89] ACER, 2017, p. 7.

One constraint experienced by ACER has been not gaining permission to view the actual test items of certain cross-national programmes. Depending on the kinds of assurances the developers of the reporting scale provide regarding the security of certain items, and the willingness of programmes to share test items, this constraint could conceivably interfere with the further development of the reporting scale.

The details of ACER's work to date are not all publicly available, but an illustrative reporting scale is put forward by ACER in their report. This tool could facilitate flexibility in terms of the grade location of the three SDG education levels. The scale is designed as a continuum specifying, for instance, a progression of competencies which would need to be attained between Grades 3 and 6, meaning one would be able to use the scale to set proficiency levels for Grades 4 and 5. This is obviously important in the context of dissimilar school system structures described in Section 2.4.

For a second phase of work referred to as validating the scales, the ACER report recommends that existing items used for the reporting scale be combined in new tests which would be run in small samples of students across a small but diverse set of countries. This would allow for the presentation and ordering of the competency descriptions in the reporting scale to be fine-tuned. In fact, it is suggested that differences in the learning processes of different countries could result in there being more than one reporting scale for each learning area.

The obvious question is the extent to which the validation exercise, which takes time (30 months, according to ACER) and carries costs, would change the original reporting scale constructed simply on the basis of expert pedagogical opinion and already available data. The answer to this question could help to justify (or fail to justify) the investment in the validation process, and would help to determine whether the UIS could begin to ask countries to begin using the scale without this validation phase. Of course, it would be difficult to answer this question about the value added by the validation exercise without first running this validation. However, informed speculation would be better than no cost-benefit consideration at all.

The MSI proposal refers to proficiency scales, as opposed to reporting scales, but essentially the purpose is the same. In MSI's approach, there would be three proficiency scales per learning area, focussing on competencies required by Grades 3, 6 and 9. ACER's approach of a continuum cross all three SDG levels is thus not followed. Moreover, the proficiency scales envisaged by MSI would express competencies largely in terms of more general performance level descriptors, as opposed to ACER's more specific descriptions, which are in effect descriptions of typical items. For instance, a performance level descriptor for Grade 6 language envisaged by MSI states that a student should "demonstrate comprehension … by using textual evidence to summarise and/or analyse a text". Or the student should use "context and word structure to determine the meanings of words".[90] The proficiency scale would then be used by experts to find corresponding cut scores in cross-national or national programmes. At this stage, test items would be considered by the experts, though no statistical analysis would occur. As in the ACER proposal, established procedures to combine the judgements of the experts would be employed.

Even if one excludes ACER's second validation phase, MSI's proposal involves lower production costs in arriving at the tools. However, less detail in the tools envisaged by MSI, and arguably less equivalence in the proficiency cut scores of the different national and cross-national programmes which would result, are matters which would need to be considered. Very importantly, the work that would have to occur across all

[90] UIS, 2017l, p. 8.

or most countries in determining cut scores in national assessments would be rather different, depending on whether the tools looked more like ACER's reporting scales or MSI's proficiency scales. Of course, some kind of hybrid between the two is possible. The resultant equivalence of the cut scores is one matter to consider. But the ease of use of the tools is another vital matter. The greater specificity of ACER's reporting scale, and above all the fact that it deals with more than three grades, could be seen as beneficial for country-level implementers of the tool. On the other hand, a more generic tool that simply required users to compare national systems against global performance level descriptors could be considered easier to implement, without large amounts of capacity building. The question is whether creating pressure for more capacity building is not a good thing.

As already mentioned, both ACER and MSI recommend the use of established procedures for a group of pedagogical experts to determine cut scores. These procedures are relatively familiar to experts in the United States and other developing countries. However, one can expect challenges when introducing these methods into the various national contexts, with differing assessment traditions and different linguistic realities.

Neither of the two proposals discusses a crucial matter in much depth: difficulties arising out of the fact that different countries and categories of countries may have very different understandings of what constitutes minimum in the minimum proficiency level of Indicator 4.1.1. The MSI proposal implicitly accepts that if, say, PIRLS and SACMEQ documents display different notions of minimum, then the two should be accepted as equivalent, even if they are not. This explains the term social moderation – it is accepted that differing expectations in different societies should to a degree be tolerated in the international reporting systems. Moreover, complexities relating to the relationship between education level and typical proficiency statistics are not discussed. The question of whether one plans for proficiency benchmarks that produce higher proficiency statistics at higher grades – the outcome of practices in LLECE and PASEC *(see Section 2.4)* – or perhaps even the reverse, remains an important matter that should be considered explicitly.

Neither proposal addresses specifically the question of the comparability of national proficiency statistics over time. It seems as if the reporting (or proficiency) scale could in fact facilitate comparability over time within a country, even if comparability across countries is weak. Experts in different countries may interpret and utilise the reporting scale differently, but if national teams of experts remained fairly consistent over time, it is likely that these teams would use the reporting scale consistently across years. However, this benefit is more likely to be realised if clear and explicit guidance is provided.

One point illustrates the kind of guidance that would be needed. What would be undesirable is considerable effort after one run of a national assessment to determine a cut score, using the reporting scale, and in subsequent runs to then use the same cut score if the national assessment is not designed to produce highly comparable results over time. In such a situation, re-applying the reporting scale after each run of the testing programme would be necessary.

Would the reporting scale have to be revised in future? Would ACER's validation phase have to be repeated? In other words, what are some of the sustainability questions around the proposed tools? These questions should be explored further, but one can speculate that as learning processes are not static – over time the grades at which specific competencies are acquired could move around – and as items in the cross-national assessments expire and become publicly available, the need for a refreshed reporting scale would arise. However, one can speculate that such a refreshing would only become necessary every ten or so years.

Crucially, the more specificity in the tool in terms of competency descriptions, the greater the need for periodic updates is likely to be.

Strategy 2 of Treviño and Órdenes overlaps to a large degree with the proposal described above.

4.3 Recalibration through the running of parallel tests (Rosetta Stone)

A three-page proposal by the IEA outlines the Rosetta Stone solution.[91] This solution deals only with the primary level. The proposal states that in the next waves of five regional assessment programmes – the same five described in Table 3 – sub-samples of students in three to five countries per programme would write not just the regional tests, but also Grade 4 TIMSS and PIRLS tests. This would produce a concordance table allowing for the conversion of, say, PASEC scores to TIMSS or PIRLS scores. This, in turn, would permit the calculation of Indicator 4.1.1 proficiency statistics, using the international benchmarks in TIMSS and PIRLS. How one might deal with possible complexities arising from the fact that the five programmes do not test Grade 4, but Grade 6 (or Grade 5, in the case of PASEC) is not discussed. The proposal states that the resulting statistics would be used for Indicator 4.1.1(b), or the end of primary education.

What is not discussed is how the Rosetta Stone might have to be updated where the five regional assessments experience explicit or suspected shifts in their standards (something which appears to have occurred in the past – see Section 2.5).

Strategies 3, and to some extent Strategy 4, of Treviño and Órdenes, display similarities with the Rosetta Stone proposal. Their Strategy 3 involves having test items which are shared across all the cross-national tests at similar education levels in order to make cut scores equivalent psychometrically. Their Strategy 4 is the ideal, certainly difficult and costly to achieve, of having a global assessment system using equivalent instruments across all countries.

5 Framework for assessing the costs and benefits

The analysis in this section follows three key steps. Firstly, a table takes stock of the types of assessment each of the three proposals described in Section 4 would rely on. This is to help gauge the potential global scope of the three proposals, in the context of the statistics and figures presented in Section 3. Secondly, the costs and benefits of relying on each of the five assessment types for Indicator 4.1.1 are evaluated, using a second table. Thirdly, the costs and benefits of the three proposals are discussed, given in part what emerges from the two tables.

Table 6 is relatively straightforward. The first proposal (statistical recalibration of existing data) is limited to using data from cross-national programmes, as it relies on the common standards within each of the programmes, and country-specific overlaps between them, to produce a global and harmonised dataset. The second proposal (social moderation) involves focussing initially on a core set of countries, namely those in the cross-national programmes, in order to produce a reporting scale, which would be the outcome of work based on document analysis, expert opinion and possibly new data analysis. However, once completed, the reporting scale could be used by anyone to identify roughly comparable proficiency benchmarks within national assessment programmes and even examinations. The Rosetta Stone proposal deals only with the

[91] IEA, 2017.

primary level, and only with TIMSS, PIRLS and five regional programmes. However, it is conceivable that the Rosetta Stone's linking assessments could be run in parallel with national sample-based assessments, in particular those of large countries so that, for instance, China would be able to convert scores from its new national system to the TIMSS or PIRLS scales.

The figures from Section 3 indicate that the first and third proposals are seriously limited in their scope, relative to the second one, unless more countries join the cross-national programmes. Specifically, not using national programmes, as in the first and third proposals, roughly halves the world population covered at the primary level *(see Table 4)* and puts countries covered at around 119, as opposed to around 155 if national assessments are included *(see Table 2).*

Table 6. Relationship between proposals and assessment types

	Statistical recalibration of existing data	**Pedagogically-informed determination of cut scores (social moderation)**	**Recalibration through the running of parallel tests (Rosetta Stone)**
PISA, TIMSS and PIRLS	Used	The core	Would be used
Regional cross-national assessments	Used	The core	Would be used
Sample-based national assessments	Cannot be used	Can be used	Mostly cannot be used though perhaps feasible for large countries
Censal assessments	Cannot be used	Can be used	Would not be used
National examinations	Cannot be used	Can be used	Would not be used

The next table draws from the critical issues discussion of Section 2 in constructing a picture of the costs and benefits of using the five assessment types for Indicator 4.1.1. This will facilitate conclusions in relation to the three proposals, as well as the suggestions for the way forward appearing in Section 6.

Table 7. Costs, benefits and assessment types

	PISA, TIMSS and PIRLS	Regional cross-national assessments	Sample-based national assessments	Censal assessments	National examinations
COSTS					
Financial cost for countries	Relatively high	Not too high	Would vary by country and strategy chosen. Not making use of the economies of scale of cross-national systems could push costs up. But less international transacting and travel could reduce costs. Donor funding can reduce national costs, but such funding can be unpredictable.	High	Given that examinations have a long history, costs would not be seen as new.
Financial cost for UIS	Low. Some investment in analysis and desktop equating of results is likely.	A bit higher than for the previous column, given weaknesses such as insufficient technical documentation.	The cost of developing new manuals, documenting country cases, providing feedback to countries and capacity building would be relatively high. This is assuming the UIS pursued these activities and did not simply accept what countries submitted.	Costs would be even higher than in the previous column as censal systems do not really follow universally established methodologies.	Costs would be even higher than for the previous column given that support would have to be provided for particularly innovative and very country-specific data work.
BENEFITS					
World coverage (see Table 1*)*	At the primary level, around a quarter of the population-weighted world. Higher at the lower secondary level.	At the primary level, around 16% of the population-weighted world (though higher when SEA-PLN is taken into account). Nothing at the secondary level.	At the primary level, around two-thirds of the population-weighted world. A bit lower at the secondary level. Due to data gaps, figures are under-estimated.	At the primary level, only 12% of the population-weighted world. A bit lower at the secondary level.	As for the previous column, but here coverage statistics are especially likely to under-estimate realities.
Comparability across countries	High within each programme, relatively easy to equate across programmes insofar as technical documentation is comprehensive and there are many doubloon countries.	Almost as high within each programme, less easy to equate across programmes. Differences across programmes in the selected grade complicates comparisons.	Low due to a large variety of sampled populations, different methodologies, possible interference by some governments.	Even lower degree of comparability given fundamentally different measurement approaches across countries.	As for previous column, but even more serious comparability difficulties.

	PISA, TIMSS and PIRLS	Regional cross-national assessments	Sample-based national assessments	Censal assessments	National examinations
Comparability over time	Mostly high. Problems described in Section 2.5 are relatively easy to resolve. Independence from national authorities and a global pool of experts all facilitate comparability over time.	As for previous column, with some limitations, in particular due to a smaller and less experienced pool of experts.	This would be as high as for the cross-national programmes in those countries following rigorous methods and where assessors are relatively independent of government.	The non-sample nature of the assessment makes maintaining secure and repeated items difficult, making comparability over time difficult to achieve technically.	The lowest of all, given the impossibility of maintaining secure anchor items.
Timeliness of the statistics	One year lag, so high timeliness.	Varies – one to four years.	Would vary by country, but likely to be higher than for the regional programmes, where single countries easily delay the entire process.	Probably best timeliness, given the fact that these systems are mostly linked to high-stakes school accountability.	Lowest lags, given the fact that examinations are linked to providing students with qualifications.
Scope for public buy-in and policy impacts	The fact that the assessments are seen as fair and independent and the fact that they allow for international comparisons, make the results highly influential. In part ideologically driven opposition from teacher unions poses a risk, but the same can be said for all assessment types.	Largely as for the previous column, though concerns around the accuracy of the statistics, and the transparency of methods used, are more common.	Assuming that the national assessment is conducted rigorously, statistics are likely to be respected by country experts, though the public may trust improvements seen in a national system less than improvements seen in a cross-national system. The fact that the assessment is national, improves the chances that results will positively influence curriculum design and teacher training. A poorly conducted national assessment may not be taken seriously, and if it is, there is a risk that weak information will inform policy.	Given that technically censal assessments present serious challenges, and given that these assessments tend to carry high stakes and be linked to school-level accountability, buy-in can easily be undermined by doubts around technical rigour and concerns around the fairness of the accountability arrangements.	Indicator 4.1.1 proficiency cut-offs are likely to be lower than cut-offs which have historically been used to determine for instance promotion to the next grade. This contradiction could be seen as confusing by the public, reducing the chances of a clear policy impact.
Scope for national capacity building	Limited, particularly in more technical areas such as item development, test design and data analysis.	Largely as for previous column, but the regional nature of the programme would increase the chances of country experts being directly involved in the more technical aspects.	If country experts have access to good materials and training programmes, national assessments can play a large role in building capacity at the national level.	As for previous column, but here the capacity required would be even wider, covering areas such as the measurement of socio-economic contexts to ensure the fairness of school-level accountability processes.	Insofar as assessment experts would simply apply traditional approaches in using, say, examinations to report on Indicator 4.1.1, there would be limited scope for capacity building.

In the light of all the preceding discussions in this report, the following conclusions can be made about the three proposals.

Statistical recalibration of existing data

Of the three proposals, it is easiest to draw conclusions about this one. There are three key reasons for not adopting this proposal as the official source for Indicator 4.1.1. Firstly, the fact that the final linking model adopted was in part the result of expert opinion, and the fact that alternative models would produce different values,[92] would make it difficult to convince countries to accept the results. Countries would ask whether the decision not to use an alternative model, one yielding a higher value, was not used. Of course, many widely accepted statistics are based on methods where analysts took somewhat subjective decisions around choosing which method to use. The use of different item response theory methods in different assessment systems is a case in point. However, achieving some kind of international consensus around the optimal methods to use in this first proposal would probably be difficult.

Secondly, in future years, as the Altinok datasets are updated as new assessments are run, not only would new values be added to old values, the old values would in many instances have to change. To illustrate, if new cases of Altinok's doubloon countries, or countries participating in more than one international assessment, emerged, this could improve the linking process and make it necessary to revise past values. Not revising past values would mean that trends would be less accurate than they should be. While some economic indicator values, such as GDP, experience adjustments of past values, and such adjustments can bring about better monitoring, this is not a common practice in the UIS system and countries are unlikely to accept fluid statistics on Indicator 4.1.1.

Thirdly, the coverage of this proposal would be unacceptably low, at least until more countries joined the cross-national programmes.

One could add a fourth disadvantage, but this disadvantage applies to all the three proposals: problems arising from comparisons across different grades, when we know that grade does, to some extent, systematically influence proficiency *(see Section 2.4).*

However, there are good reasons for UIS or some other organisation to invest in periodic analyses of the type produced by Altinok. This type of analysis provides an excellent reality check for the UIS and individual countries evaluating proficiency statistics emerging from national assessments (which, as is argued below, should play an integral part of the whole Indicator 4.1.1 reporting system). Even countries not included in Altinok's dataset could use the dataset's values on other, similar countries as a guide to what proficiency statistics to expect.

Pedagogically informed determination of cut scores (social moderation)

As has been argued in Section 4.2, the reporting scale at the centre of this proposal seems necessary in any future scenario. In this sense the second proposal is strong. The reporting scale could constitute a major innovation in the education planning of many countries. As with any innovation, there would have to be

[92] Altinok, 2017, p. 47.

iterative cycles of development, piloting, refinement, implementation and then re-development. As has been discussed above, it is difficult to evaluate at this point whether validation through the running of new assessments specifically designed to validate the scale must occur before countries can begin using the reporting scale to determine cut-offs in national assessments. As outlined in Section 6, it is assumed that use without assessment-based testing of the scale is possible.

A further strength with the second proposal is that it would considerably expand the coverage of the data collection system, relative to a scenario where only cross-national assessments were counted. For instance, coverage at the primary level would double, in terms of the population-weighted world, if sample-based national assessments were included. As indicated in Table 7, bringing national assessments into the UIS reporting system would realise benefits such as more widespread capacity building, and an improved ability to solve national problems. This would, however, depend in part on the level of investment by the UIS and other development agencies in capacity building.

In Section 6, it is assumed that a tool similar to ACER's reporting scale, as opposed to the MSI's proficiency scale, would be adopted. There are two advantages with the former. The benefits of a continuum that covers all grades, and not just three grades, has been discussed. In fact, Indicator 4.1.1 is itself open-ended when it comes to determining the grade of students assessed. A second apparent advantage, which has not been discussed, is that a system which explains in fairly detailed terms the skills to be attained by students, is probably more likely to be seen as credible. The general public may in fact be sceptical about proficiency benchmarks expressed in broad terms, and wonder why specific skills, expressed almost in terms of typical questions, are not made clearer.

Recalibration by running parallel tests (Rosetta Stone)

This proposal, like the first one, is limited in terms of world coverage, because so many countries do not participate in cross-national assessments. Moreover, the proposal focusses just on the primary level. There are no financial cost figures available for this proposal, though one can assume the costs would be substantial. Moreover, complex decisions may have to be made on how the five regional programmes, or several national governments, would be compensated for including the burden of running additional tests.

Like the first proposal, the Rosetta Stone solution appears valuable as one component of the future Indicator 4.1.1 reporting system, though it would be inadequate as the core of the system, largely due to coverage problems. Importantly, this proposal in many ways attempts to do what the validation phase of ACER's proposal would do. Both would enhance the comparability of scores and proficiency statistics across the cross-national programmes focussing on the primary level. But they would do this in different ways. While ACER's validation proposal is item-focussed, and specifically aimed at improving the sequencing of proficiency descriptions along the reporting scale, the Rosetta Stone solution is more test-focussed and aims to equate scores across different programmes. The Rosetta Stone solution has one important advantage relative to the validation phase proposal: it does not require the sharing of secure items by programmes. Rather, the IEA would retain relatively strong controls over the security of the items included in the separate test. This is an important factor given that programmes are often reluctant to share secure items with each other, as they can lose some certainty over how secure those items are.

6 Possible ways forward

This section draws on the previous sections in outlining a possible way forward. In some instances, various pathways are described. A way forward may seem affordable, politically acceptable and technically possible However, it nevertheless may have risks, some of which are pointed out. The way forward described here is intended to highlight how various factors can influence the difficult task of reporting on Indicator 4.1.1.

The way forward is organised in terms of seven specific outputs, numbered 'a' to 'g'.

a. Release of the first version of the UIS reporting scale

Work on the reporting scale has already begun, and it appears this work should continue, no matter how reporting on Indicator 4.1.1 occurs. A key output would be a widely publicised version of a reporting scale bearing official UIS endorsement. It would come with a clear description of the methodology behind it, and a guide on how national experts would use it to determine proficiency benchmarks that are applicable in a national assessment, in other words, cut scores which would be used to determine proficiency statistics for Indicator 4.1.1 that are reported to UIS. The guide would explain how national experts should use a combination of qualitative methods, such as evaluation of test items, and quantitative methods, involving the analysis of raw item-level data, to identify their proficiency benchmarks.

Concerns about across-country comparability would be addressed in the guide. The guide would address the fact that, for various reasons, proficiency statistics are not be strictly comparable across countries. However, they should be sufficiently comparable, and some description of what this means would be explained. One of the factors that limits across-country comparability is the fact that, to some extent, one would be comparing different grades. A further factor would be that each country would be free to determine its cut scores within a range. The reporting scale would thus have a range along on the scale within which a country could determine its cut score for, say, Grade 2. In fact, ACER's illustrative reporting scale does have ranges per grade, and the grade-specific ranges are overlapping. Legitimate reasons for a country to choose a low proficiency benchmark for a specific grade would include the fact that educational quality was under-developed in the country and that it was necessary to use a realistic benchmark.

The scores of the cross-national programmes would appear in the reporting scale. So, for instance, the reporting scale for reading would have PIRLS, SACMEQ, LLECE and other scales attached. Moreover, there would be international proficiency benchmarks for all grades covered by the cross-national programmes. For instance, there would be a Grade 6 cut score because SACMEQ and LLECE cover this grade. In this case, a fairly rigorous across-programme comparability should be pursued. There would thus be one cut score for Grade 6, expressed in terms of a SACMEQ score, the equivalent LLECE score, and the central metric of the reporting scale itself. In ACER's illustrative reporting scale, this central metric reflects a value of around 125 for Grade 6. In reaching these equivalences, the kind of linking done by Altinok for doubloon countries should be a part of the analysis, though ACER's proposed item-focussed analysis should also play a role.

Should more than one proficiency benchmark cut score per cross-national assessment be possible? Should there be a low benchmark for developing countries, and a high one for developed countries, along the lines of Altinok's two benchmarks? This is debatable. A clear disadvantage with this is that one would then have to make somewhat arbitrary decisions about which of the two thresholds to apply when choosing one, official, proficiency statistic for a specific country. We can assume that some official Indicator 4.1.1 indicator

values will come from the cross-national programmes. If there are two benchmarks, then where does one draw the line between developed and developing countries? What seems better than attempting to draw such a line is to accept a single relatively low proficiency benchmark that is then applied for a specific grade to all relevant countries and programmes. The implication of this is that there would be one proficiency benchmark for Grade 6 which would be applicable to students in SACMEQ and LLECE, but then another benchmark for Grade 5 applicable to students in PASEC, another one for Grade 4 applicable to students in TIMSS and PIRLS, and so on.

The recommendation here is that the full range of each cross-national programme's scores be reflected on the reporting scale. Thus, in the case of TIMSS Grade 4 mathematics there would be a specific TIMSS score located within Grade 4 on the reporting scale (this would be the TIMSS proficiency benchmark), but then scores above and below this point would be reflected as, say, the Grade 5 and even Grade 6 proficiency levels. These extensions above and below the main grade-specific proficiency benchmark could be useful background information, though they would not serve any specific reporting purpose for Indicator 4.1.1. No-one would use a TIMSS score to determine a proficiency statistic for Grade 5 (except perhaps in the few cases where TIMSS Grade 4 tests are in fact applied in Grade 5). Of course, one could simplify all this by specifying only the single score corresponding to the relevant proficiency benchmark on the reporting scale, without any upward or downward extensions.

To be clear, the above discussion implies that the existing programme-specific competency thresholds would not be used for Indicator 4.1.1 purposes, and would not be reflected in the UIS reporting scale. They could inform where to place new grade-specific benchmarks which were sufficiently accommodating for developing countries. But ultimately, different programmes focussing on the same grade would be subject to equivalent benchmarks on the reporting scale.

The reporting scale itself would not refer to the (a), (b) and (c) education levels of Indicator 4.1.1. This complexity would be dealt with separately and is discussed below.

Finally, what is said here implies that a first version of the reporting scale would be released for use by countries without running new assessments to validate the reporting scale. Running such a validation phase would be ideal, but the cost and above all the delays that this would bring about seem concerning.

In how many languages would the reporting scale be published? This question is not easy to answer, but it is an important one. The precedent is that the UIS, a relatively technical body, publishes mostly in English though, crucially, the main UIS annual questionnaires for countries are also available in French. UNESCO, which is more political, publishes key reports, in particular the large Global Education Monitoring Report, in all six official UN languages. It is perhaps best to publish the full reporting scale report in English, with summarised versions available in all six official UN languages.

One aspect of the reporting scale that would need to be updated fairly frequently, is the correspondence of the scores on the cross-national programmes. New programmes would have to be included as soon as their data became available. Moreover, any shift in programme standards would lead to an adjustment in the reporting scale.

b. Release of new proficiency statistics drawing from cross-national programmes and the reporting scale

This would be a relatively easy output to produce, once the grade-specific benchmarks applicable to cross-national assessment data had been inserted in the reporting scale.

The new proficiency statistics would be similar to the values already published through UIS.Stat (those values reflect proficiency benchmarks determined within each of the cross-national programmes). The new statistics should probably replace the existing Indicator 4.1.1 statistics on UIS.Stat, though the previously released values should be available online as a backup. Some analysts may already have used these previously released values and should have continued access to them. In part, the new statistics would update the datasheets through the insertion of post-2015 values, for instance those of PIRLS 2016 (but even SACMEQ 2013, which has not been available previously).

It does seem important for the UIS to promote the use of its reporting scale to compute official Indicator 4.1.1 statistics where the source for the data is a cross-national programme. Not doing so might weaken the understanding and acceptance of the reporting scale in countries analysing data from their own national statistics.

The statistics referred to here would be based on a linking of grade to education level that would be undertaken, but also transparently justified, by the UIS. It seems optimal to follow Altinok's approach of counting Grade 4 results (and anything below that) as corresponding to the lower primary level, but some kind of justification would have to be put forward.

The international database referred to here would ideally be updated every time there was a new set of data produced by a cross-national assessment programme.

c. Country access to the UIS questionnaires for Indicator 4.1.1

A request for Indicator 4.1.1 information would go out to countries through the existing UIS system of country questionnaires. A separate questionnaire would be designed for Indicator 4.1.1, and the methodology for this questionnaire would be somewhat different to that of the existing questionnaires. Above all, given that Indicator 4.1.1 statistics would inevitably be less comparable across countries than virtually all other national statistics published by the UIS, considerable attention would be placed on what was being measured and how within each country. Moreover, to inform capacity building, there would be considerable focus on national plans to improve the monitoring of learning outcomes, and how each country views its current capacity constraints.

The Indicator 4.1.1 questionnaire would be separate, in part because those in the national education authority filling it in are likely to be different from the people who fill in the other UIS questionnaires dealing with, for instance, enrolments and expenditure. It would be made clear that the questionnaire was focussing mainly on sample-based national assessments, though there would be some questions on the country's participation in cross-national assessments. Figures from Section 3 have indicated that at least 62 countries, covering around two-thirds of the world's population, would be in a position to provide information on sample-based national assessments at some education level. These figures are under-estimates insofar as their sources clearly exclude data on some countries.

The questionnaire should reassure respondents that UIS understands fully that comparisons across countries based on the submitted data would be difficult (due to various reasons explained in the current report and elsewhere), and that the UIS will display sensitivity to this in the international reporting process. It should also be made clear that the UIS recognises that many countries may have underdeveloped systems that impact their capacity to collect quality data. These reassurances are necessary to avoid a situation where countries may be reluctant to submit data because their systems are weak, or because they do not see the point of submitting statistics based on a national system that is very different to the systems of other countries, into an international reporting system, perhaps because they believe resultant rankings would be unfair.

Should censal national assessments and examinations also be covered by the Indicator 4.1.1 questionnaire (see the discussion in Section 2.6 in relation to differences between sample-based and censal national assessments)? To avoid complexities, and because the world coverage gained by including examinations is limited, examinations should probably not be included, though this is debatable. Censal national assessments should probably be within the scope of the questionnaire, even if there is a considerable risk that these assessments are less effective at gauging trends over time relative to sample-based national assessments.

Figure 8 and its sources suggest that including censal national assessments would be especially important for the end of primary (see for instance the countries in Europe marked in orange).

The UIS questionnaire on Indicator 4.1.1 would come with a description of the UIS's work on a few country-focussed case studies. These case studies would involve examining how the selected countries were completing the UIS questionnaire, using the reporting scale, and what technical assistance each country needs most. It would be explained that the studies would, to some extent, be conducted by researchers who are independent of the UIS (and of the country governments involved) in order to facilitate constructive criticism of the whole process.

The UIS questionnaire would lead respondents to determine the official statistics for the six main indicators, Indicator 4.1.1 (a), (b) and (c) for the two learning areas, as well as the breakdowns of the six values by gender. Obviously, the ability of a country to do this would depend on the existence of the necessary national assessments. There would be some guidance and parameters relating to the correspondence between grade and the three education levels, but within these parameters countries would have some leeway in determining whether, say, a Grade 4 assessment should be counted under (a) or (b). However, the questionnaire would require the country to state its reasons for any choices made. One important statistic that should be provided by each country would be the average age of the assessed children, and the point in the year at which this average age was measured. In fact, going further and providing a breakdown of

children by age would be ideal. It would be made clear in the instructions accompanying the questionnaire that the leeway given to countries in attaching grades to SDG education levels would be yet another reason why the UIS would exercise much caution when comparing countries.

A series of questions would aim to gather background information on the national assessment, information which would help the UIS and others interpret the meaning of the official statistics. The topics to be covered by these questions should be obvious: the assessment framework used, item development, test construction, test administration processes, data processing, and scoring. A separate set of questions would cover the processes followed by the country in using the UIS reporting scale, and its accompanying instructions, to determine SDG proficiency benchmarks within the scores of the national assessment. The questions relating to the processes of the national assessment could be considered sensitive by some countries. These processes may not be ideal and certain countries may consequently decide not to publicise some information. The questionnaire should be constructed in such a way that it does not encourage evasive or misleading responses. This can be achieved by explicitly allowing a response such as 'Information cannot be submitted due its sensitivity'.

It would be made clear to countries that the comments inserted in the questionnaire would not be made available to people outside the UIS. However, countries would be urged to produce one summary paragraph per national assessment and grade which it believes should be made public so that the submitted SDG 4 statistics would be properly interpreted.

Crucially, the questionnaire should encourage countries to make technical documentation and analyses resulting from the national assessment available to the UIS. This is perhaps best achieved through a password-protected uploading portal. Two options should be possible. Documents the country believes can be available for anyone should be marked as such. There should be virtually no limit to what a country includes in this set. Secondly, documents the country would like to share with UIS experts, but which the UIS agrees would not be shared beyond these experts, or beyond the UIS, would be marked as belonging to a secure category. The end result would be an online repository where documents in the first category would be downloadable by anyone, and documents in the second category would be listed, but not downloadable. An e-mail address of a national official would be specified through whom access to a document may be granted, depending on the purpose of the request. The online library described here would not be costly for the UIS to set up, yet it has the potential to vastly improve information-sharing and learning among countries. Importantly, the UIS would not make any judgements or assurances, regarding the technical rigour of the shared documents. This library would serve a purpose somewhat similar to that of the IIEP's Polipolis repository, which shares education policy documents.

Some questions in the questionnaire would deal with cross-national assessments (existing UIS questionnaires already gather basic information on participation in these assessments and the associated capacity building needs[93]). One important question here would elicit a country's preference with respect to the linking of grade to education level in the case of cross-national assessments. For instance, countries participating in TIMSS Grade 4 testing would specify whether they thought it was more appropriate for these results to correspond to lower primary or the end of primary – or countries could state that they were unsure about how to draw this link. It would be made clear that this information was being collected to inform the

[93] http://uis.unesco.org/en/methodology, under heading 'UIS Questionnaires' (accessed June 2018).

UIS's work, but that the UIS would, in its use of cross-national assessment data, use the same grade-level link across all countries (as has been indicated above).

One thing that should be considered is the addition of a few open-ended questions relating to the country's satisfaction with the cross-national assessments in which it participates. This could take the form of an initial rating of various aspects of each programme – perhaps the reasonableness of budget costs, data collection strategies, technical rigour and transparency of data processing work – followed by space for comments.

Importantly, countries would be encouraged to provide information on national assessments and to use them to derive statistics for Indicator 4.1.1 even if cross-national assessments appear to cover specific education levels. As discussed below, it is proposed that statistics from the cross-national assessments and statistics from national assessments be kept relatively separate in the UIS reporting systems, though at some point the two have to come together.

The processes to collect information on national systems that are outlined in the preceding paragraphs differ from existing UIS data collection processes. The online documents repository would be an important innovation. This shift could be considered useful not just for Indicator 4.1.1, but for the UIS data collection processes in general. For instance, the repository could be expanded to include documents on other education themes, such as teachers, enrolment and education financing.

Constructive criticism of the questionnaire, and findings from the case studies, would enable the UIS to implement improvements based on country needs. It would also be explained that future versions of the questionnaire would probe in more depth the comparability over time of the results from the national assessment systems.

d. Release of country-specific Indicator 4.1.1 statistics on UIS.Stat based on the new survey

The SDG 4 statistics collected through the questionnaires would be published online by the UIS. These would be kept completely separate from the SDG 4 statistics based on cross-national programmes (those statistics were discussed above under 'b'). The statistics drawing from the questionnaires should be easily linked within the online system to the summary paragraphs which countries said should accompany the statistics. The UIS.Stat online system would display a prominent message to users indicating the serious limitations of the statistics with respect to comparability, in particular across countries, but possibly even over time.

The UIS would only filter out obviously nonsensical statistics. The usual processes followed in the past of returning to countries which clearly had problems understanding or filling out questionnaires would be followed before the release of the statistics.

e. Release of an evaluation of the submitted values by the UIS

Following the release of the questionnaire-based statistics (see 'd'), the UIS would release a report that evaluated the statistics through an analysis of the internal coherence of the dataset, and comparisons against other statistics, in particular those emerging from cross-national assessments. This report would be relatively technical, and not an SDG progress report directed at a large readership. Its main audience would be national education authorities and organizations assisting these authorities. It would draw extensively from the comments collected in the questionnaires, without sharing information on individual countries that

would reflect poorly on national administrations, and which could reduce cooperation in future runs of the questionnaire. In other words, the focus would be on general patterns, including those relating to problems within national systems, and exemplary practices observed in specific countries.

The emphasis would be on evaluating the degree to which national assessments are able to provide a basis for country-specific progress, which, subject to important provisos, would mean that the world was better able to measure progress. Comparisons across countries, and the calculation of regional and global aggregate proficiency statistics, would be done with due caution. For instance, one could emphasise that such aggregate statistics were approximate, or one could provide statistical ranges. The attachment of grades to the three SDG education levels would also have to be done with much caution. The preferences of countries regarding this link would have to be clear. It should be clear where one was mixing, say, Grade 2 and Grade 4 results within the same analysis.

f. Release of country case studies

This output would deal with issues similar to the evaluation report referred to under 'e'. It should probably be maintained as a separate output, however, as it would be produced through a separate process, and delays in one output should not delay the release of the other. However, preliminary findings from 'f' should feed into 'e'. Reference has already been made to the country case studies under 'c' above. The case studies should be designed in such a way that they have built-in processes dealing with likely differences of interpretation between, in particular, the researchers and national authorities. The reports should not ignore gaps and challenges in the national systems and processes, but criticism should be constructive. Very importantly, for a country to participate in the case studies, it should be a condition that raw data be shared with the researchers, if necessary subject to stringent controls to prevent leaks. It is probably not effective for the case studies to draw only on interviews and documents, without also involving some analysis of patterns in the data by the researchers. In the absence of some examination of the data, key aspects of the national assessment programme could be missed.

g. Official Indicator 4.1.1 statistics per country and for the world

In the above proposals, statistics from national and cross-national systems are kept separate. This separation is necessary as the two sources are so different. However, the public need for the correct or at least the best Indicator 4.1.1 value per country must be acknowledged. It is probably not acceptable for, say, UNESCO's Global Education Monitoring Report to provide two separate tables corresponding to the two sources, even if many would argue that this would be the most transparent and methodologically correct approach. Assuming that a choice must be made between the two, the following is proposed.

Statistics from cross-national programmes would be considered preferable, so in the case of two candidates being available, one from cross-national and one from the national system, the cross-national statistic would be used. The country's preference for linking grade to SDG education level, as far as the national assessment was concerned, would be respected. In any final table of statistics, symbols would indicate which of the two sources had been used. Moreover, there would have to be a reference to some separate source which would indicate which grade had been used for each statistic – this would be especially important in the case of statistics derived from national assessments. Importantly, any comparison over time, even an implicit comparison in the sense of two statistics from different years appearing side by side, should draw from just one of the two types of sources. In fact, it is debatable whether comparisons over time for a country, whether

explicit or implicit, should reflect more than one cross-national source. For example, should Chile's proficiency statistics for lower primary drawn from PIRLS Grade 4 and LLECE Grade 3 appear side by side? The statistics would have both been derived through comparison against the same reporting scale, but using different grades in that comparison process. There are not many such possible clashes, but the ones that exist should be carefully considered.

There is of course an important disadvantage with the proposal put forward in the previous paragraph. It may clash with the principle that countries should have control over which of their statistics are published by UNESCO. This principle is at least implicit with respect to many statistics published by the UIS, but not all. Only enrolment figures supplied by national education authorities are published by the UIS. However, enrolment ratios use UN Population Division figures which are often different from the official population estimates of countries. As mentioned in Section 1, proficiency statistics derived from cross-national assessments, over which countries have limited control, have already been published through UIS.Stat, apparently without any objections being raised by national authorities. If the principle of national determination with respect to proficiency statistics were to be upheld, then clearly the preferences of countries – in terms of national versus cross-national – would need to be collected through the UIS questionnaire system.

References

ACER (2017). *UIS Reporting Scales (UIS-RS).* Camberwell: Australian Council for Educational Research. Available from: http://gaml.uis.unesco.org/files/meeting4/UIS-RS_Concept_Note_July2017.pdf (accessed May 2018).

Agencia de la Calidad de la Educación (2014). *Informe Técnico: SIMCE 2014.* Santiago. Available from: http://archivos.agenciaeducacion.cl/InformeTecnicoSimce_2014.pdf (accessed April 2018).

Altinok, N. (2017). *Mind the Gap: Proposal for a Standardised Measure for SDG 4–Education 2030 Agenda.* Montreal: UNESCO Institute for Statistica (UIS). Available from: http://uis.unesco.org/sites/default/files/documents/unesco-infopaper-sdg_data_gaps-01.pdf (accessed April 2018).

Altinok, N., N. Angrist and H.A. Patrinos (2018). *Global Dataset on Education Quality (1965-2015)*. Washington: World Bank. Available from: http://documents.worldbank.org/curated/en/706141516721172989/pdf/WPS8314.pdf (accessed April 2018).

Belisle, M., E. Cassity, R. Kacilala, M.T. Seniloli and T. Taoi, T. (2016). *Pacific Islands Literacy and Numeracy Assessment: Collaboration and Innovation in Reporting and Dissemination.* Paris: UNESCO. Available from: http://unesdoc.unesco.org/images/0024/002468/246812e.pdf (accessed June 2018).

Benveniste, L. (2002). "The political structuration of assessment: Negotiating state power and legitimacy". *Comparative Education Review,* 46(1), pp. 89-118.

Brookings Institute (2015). *Assessment for Learning: An International Platform to Support National Learning Assessment Systems.* Washington. Available from: https://www.brookings.edu/wp-content/uploads/2015/12/A4L-Concept-Note-Discussion-Document-12115.pdf (accessed July 2018).

Bruns, B., D. Evans and J. Luque (2012). *Achieving World Class Education in Brazil: The Next Agenda*. Washington: World Bank. Available from: https://openknowledge.worldbank.org/bitstream/handle/10986/2383/656590REPLACEM0hieving0World0Class0.pdf?sequence=1 (accessed February 2016).

Carnoy, M., T. Khavenson, I. Fonseca and L. Costa (2015). "Is Brazilian education improving? Evidence from Pisa and Saeb". *Cadernos de Pesquisa*, 45(157). Available from: http://www.scielo.br/scielo.php?pid=S0100-15742015000300450&script=sci_arttext&tlng=en (accessed May 2017).

Cheng, Y.C., K.H. Ng and M.M.C. Mok (2002). "Economic considerations in education policy making: A simplified framework". *The International Journal of Educational Management,* 16(1), pp. 18-39.

Clarke, M. (2012). *What Matters Most for Student Assessment Systems: A Framework Paper*. Washington: World Bank. Available from: https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1 (accessed October 2015).

CONFEMEN (2012). *Synthèse régionale des résultats PASEC - Document de travail*. Dakar. Available from: http://www.confemen.org/wp-content/uploads/2012/03/Synthese_PASEC_VII-VIII-IX_final.pdf (accessed December 2012).

CONFEMEN (2015). *PASEC 2014 Education System Performance in Francophone Sub-Saharan Africa*. Dakar. Available from: http://www.pasec.confemen.org/wp-content/uploads/2015/12/Rapport_Pasec2014_GB_webv2.pdf (accessed July 2018).

Department of Examinations Sri Lanka (2015). *Reviewing of Performance at Grade 5 Scholarship Examination 2015.* Colombo. Available from: https://www.doenets.lk/exam/docs/comm/Grade%2005%20-%202015%20Symposium.pdf (accessed March 2009).

Education International (2011). *Policy Paper on Education: Building the Future through Quality Education*. Brussels. Available from: http://www.ei-ie.org (accessed October 2011).

Ferrer-Esteban, G. (2013). *Rationale and Incentives for Cheating in the Standardised Tests of the Italian Assessment System*. Torino: Fondazione Giovanni Agnelli. Available from: http://www.fga.it (accessed October 2017).

Flotts, M.P., J. Manzi, D. Jiménez and A. Abarzúa (2015). *Informe de Resultados TERCE: Logros de Aprendizaje*. Santiago: UNESCO. Available from: http://unesdoc.unesco.org/images/0024/002435/243532S.pdf (accessed September 2017).

Greaney, V. and T. Kellaghan (2008). *Assessing National Achievement Levels in Education*. Washington: World Bank. Available from: http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099079993288/assessing_national_achievement_level_Edu.pdf (accessed March 2010).

Gustafsson, M. (2015). "Enrolment ratios and related puzzles in developing countries: Approaches for interrogating the data drawing from the case of South Africa". *International Journal of Educational Development,* 42, pp. 63-72.

Gustafsson, M. (2016). *Understanding Trends in High-Level Achievement in Grade 12 Mathematics and Physical Science.* Pretoria: Department of Basic Education. Available from: http://resep.sun.ac.za/wp-content/uploads/2016/06/Correcting-racial-imbalances-2016-01-25.pdf (accessed June 2016).

Gustafsson, M. and C. Nuga Deliwe (2017). *Rotten Apples or Just Apples and Pears? Understanding Patterns Consistent with Cheating in International Test Data.* Stellenbosch: Stellenbosch University. Available from: https://ideas.repec.org/p/sza/wpaper/wpapers293.html (accessed January 2018).

Hanushek, E.A. and L. Woessmann (2007). *The Role of School Improvement in Economic Development*. Washington: National Bureau of Economic Research. Available from: http://papers.nber.org/papers/w12832.pdf?new_window=1 (accessed June 2007).

IEA (2017). *IEA's Rosetta Stone: Measuring Global Progress toward the UN Sustainable Development Goal for Quality Education by Linking Regional Assessment Results to TIMSS and PIRLS International Benchmarks of Achievement*. Chestnut Hill. Available from: http://uis.unesco.org/sites/default/files/documents/Draft_proposal_for_linking_regional_assessments_to_TIMSS_and_PIRLS.pdf (accessed: June 2018).

Jerrim, J. (2013). "The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline?" *Journal of Social Policy,* 42(2), pp. 259-279.

Klein, R. (2011). *Uma Reanálise dos Resultados do Pisa: Problemas de Comparabilidade*. Rio de Janeiro: Avaliação e Políticas Públicas em Educação. Available from: http://www.scielo.br/pdf/ensaio/v19n73/02.pdf (accessed June 2018).

Makuwa, D.K. (2010). "Mixed results in achievement". *IIEP Newsletter,* XXVIII(3). Available from: http://www.iiep.unesco.org (accessed October 2010).

McEwan, P.J. (2014). "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments". *Review of Educational Research,* XX(X): 1-42. Available from: http://academics.wellesley.edu/Economics/mcewan/PDF/meta.pdf (accessed May 2014).

Ministério de Educação Brazil (2015a). *Avaliação Nacional da Alfabetização: Relatório 2013-2014.* Brasilia. Available from: http://portal.inep.gov.br/documents/186968/484421/Relat%C3%B3rio+ANA+2013-2014+-+Da+concep%C3%A7%C3%A3o+%C3%A0+realiza%C3%A7%C3%A3o/8570af6a-c76e-432a-846f-e69bbb79e4b2?version=1.3 (accessed April 2018).

Ministry of Education China (2017). "A Glance of National Assessment of Education Quality in China". [Slide presentation]. Beijing. Available from: http://uis.unesco.org/sites/default/files/documents/unesco-infopaper-sdg_data_gaps-01.pdf.

Ministry of Education Ghana (2014). *Ghana 2014 National Education Assessment: Technical Report*. Accra. Available from: https://globalreadingnetwork.net/sites/default/files/eddata/2013_NEA_Technical_Report_15May2014_wRecs.pdf (accessed April 2018).

Ministry of Education Liberia (2007). *Liberian Primary Education Recovery Program*. Monrovia. Available from: http://planipolis.iiep.unesco.org/sites/planipolis/files/ressources/liberia_plan.pdf (accessed July 2018).

Ministry of Education Liberia (2016). *Getting to Best Education Sector Plan July 2017-June 2021*. Monrovia. Available from: http://moe.gov.lr/wp-content/uploads/2017/11/Liberia-Getting-to-Best-Education-Sector-Plan-2017-2021.pdf (accessed July 2018).

OECD (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. Paris.

Pritchett, L. and A. Beatty (2012). *The Negative Consequences of Overambitious Curricula in Developing Countries*. Washington: Center for Global Development. Available from: http://www.cgdev.org/files/1426129_file_Pritchett_Beatty_Overambitious_FINAL.pdf (accessed August 2014).

South Africa: Department of Basic Education (2016). *Report on Progress in the Schooling Sector against Key Learner Performance and Attainment Indicators*. Pretoria. Available from: http://www.education.gov.za (accessed September 2016).

Tyumeneva, Y. (2013). *Disseminating and Using Student Assessment Information in Russia.* Washington: World Bank. Available from: https://openknowledge.worldbank.org/bitstream/handle/10986/16276/810910WP0100Ru0ox0379828B00PUBLIC00.pdf?sequence=1&isAllowed=y (accessed July 2018).

UNESCO (2008). *Primer Reporte: SERCE: Los Aprendizajes de los Estudiantes de América Latina y el Caribe*. Santiago: OREALC/UNESCO. Available from: http://unesdoc.unesco.org/images/0016/001606/160660s.pdf (accessed March 2010).

UNESCO (2014). *Education for All Global Monitoring Report 2013/4 – Teaching and Learning: Achieving Quality Education for All*. Paris. Available from: http://www.unesco.org (accessed May 2015).

UNESCO (2016). *Reporte Técnico: Tercer Estudio Regional Comparativo y Explicativo*. Santiago. Available from: http://www.unesco.org/new/fileadmin/MULTIMEDIA/FIELD/Santiago/pdf/Reporte-tecnico-TERCE.pdf (accessed June 2018).

UNESCO (2017). *Global Education Monitoring Report 2017/18 – Accountability in Education: Meeting our Commitments*. Paris. Available from: http://www.unesco.org (accessed April 2018).

UNESCO Institute for Statistics (UIS) (2016). *Country Readiness to Monitor SDG 4 Education Targets: Regional Survey for the Asia and Pacific Region.* Montreal. Available from: http://unesdoc.unesco.org/images/0024/002458/245829e.pdf (accessed July 2018).

UNESCO Institute for Statistics (UIS) (2017a). "Counting the Number of Children not Learning: Methodology for a Global Composite Indicator for Education". UIS Information Paper No. 47. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/ip47-counting-number-children-not-learning-methodology-2017-en.pdf (accessed April 2018).

UNESCO Institute for Statistics (UIS) (2017b). *Quick Guide no. 3: Implementing a National Learning Assessment*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/quick-guide-3-implementing-national-learning-assessment.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017c). "More than One-Half of Children and Adolescents are not Learning Worldwide". UIS Fact Sheet No. 46. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/fs46-more-than-half-children-not-learning-en-2017.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017d). "Exploring Commonalities and Differences in Regional and International Assessments". UIS Information Paper no. 48. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/ip48-exploring-commonalities-differences-regional-international-assessments-2017-en.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017e). *Metadata for the Thematic and Global Indicators for the Follow-Up and Review of SDG 4 and Education 2030*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/metadata-global-thematic-indicators-sdg4-education2030-2017-en_1.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017f). *Report of the Director on the Activities of the Institute in 2017.* Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/report-of-director-on-activities-of-the-institute-2017.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017g). *SDG Reporting: Linking to the UIS Reporting Scale through Social Moderation*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/gaml-sdg4-reporting-linking-uis-reporting-scale-social-moderation-2017-en.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017h). *The Investment Case for SDG 4 Data*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/investment-case-sdg4-data.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017i). *Global Alliance to Monitor Learning (GAML): Concept Paper*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/gaml-concept_paper-2017-en2_0.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017j). *Principles of Good Practice in Learning Assessment*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/principles-good-practice-learning-assessments-2017-en.pdf (accessed July 2018).

UNESCO Institute for Statistics (UIS) (2017k). *The Quality Factor: Strengthening National Data to Monitor Sustainable Development Goal 4.* Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/quality-factor-strengthening-national-data-2017-en.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2017l). *Constructing UIS Proficiency Scales and Linking to Assessments to Support SDG Indicator 4.1.1 Reporting*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/gaml4-constructing-uis-proficiency-scales-linking-assessments-support-sdg-indicator4.1.1-reporting.pdf (accessed July 2018).

UNESCO Institute for Statistics (UIS) (2018a). *Towards an Innovative Demand-Driven Global Strategy for Education Data*. Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/towards-innovative-demand-driven-global-strategy-education-data-2018-en.pdf (accessed May 2018).

UNESCO Institute for Statistics (UIS) (2018b). *Quick Guide No. 2: Making the Case for a Learning Assessment.* Montreal. Available from: http://uis.unesco.org/sites/default/files/documents/quick-guide2-making-case-learning-assessments-2018-en_2.pdf (accessed May 2018).

United Nations (2017). *Revised List of Global Sustainable Development Goal Indicators*. New York. Available from: https://unstats.un.org/sdgs/indicators/Official%20Revised%20List%20of%20global%20SDG%20indicators.pdf (accessed October 2017).

United Nations: Data Revolution Group (2014). *A World that Counts: Mobilising the Data Revolution for Sustainable Development*. New York. Available from: http://www.undatarevolution.org (accessed April 2016).

United States: Department of Education (2013). *An Overview of NAEP*. Washington. Available from: https://nces.ed.gov/nationsreportcard/subject/_commonobjects/pdf/2013455.pdf (accessed June 2018).

Woodhall, M. (2004). *Cost-Benefit Analysis in Educational Planning*. Paris: UNESCO-IIEP. Available from: http://unesdoc.unesco.org/images/0013/001390/139042e.pdf (accessed March 2006).