United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

SUSTAINABLE
DEVELOPMENT
GOALS

Information Paper No. 63
January 2020

Evidence-based Projections and Benchmarks for 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 specialized information.

UNESCO Institute for Statistics

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

The UIS is the official source of internationally comparable data used to monitor progress towards the Sustainable Development Goal on education and key targets related to science, culture, communication and gender equality.

This paper was written by Martin Gustafsson, Research on Socio-Economic Policy (ReSP), University of Stellenbosch. It was commissioned to inform the work of the Global Alliance to Monitor Learning (GAML).

Published in 2020 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/2020/LO/IP/63

Summary

The main focus of this is report is a model, implemented in Excel, which combines a dataset of normalised percentage proficient statistics, produced previously for the UNESCO Institute for Statistics (UIS), and demographic data, to produce projections of proficiency among children. These projections are intended to inform global and national target-setting, and the broader discussion on educational development. Specifically, the projections are intended to inform empirically discussions relating to Sustainable Development Goal (SDG) Indicator 4.1.1, which can be considered a set of 18 indicators when all dimensions relating to school level, subject and sex are considered.

The report begins by discussing, in section 2, the nature of target-setting in different non-education contexts, and how this might influence the projections, or the use of the projections. Moving forward, projections of educational proficiency that incorporate factors bringing about change seem possible, and are likely to help stimulate the kind of debate needed. Targets should be 'owned' by national authorities, who are central to the work needed to bring about educational improvement. Given vast differences across countries with respect to technical capacity, a multi-tier reporting approach, which takes into account capacity differences, and national assessment specificities, could help countries engage with SDG targets.

Section 3 explains what the Excel tool does and does not do. One thing it does not do is describe with much accuracy the baseline situations in individual countries. The global dataset compiled for the projections is the result of adjustments, which could skew the actual baseline in individual countries, as well as imputations for, at each of the three school levels, between one and two-thirds of the world's child age cohort. By extension, the tool does not set out to recommend to individual countries what their SDG 4.1.1 targets should be. What the Excel tool is designed to do is to produce baselines and projections for the world, and for the seven UNESCO world regions, which are reliable enough to inform debates around the risks and opportunities relating to the pursuit of better quality schooling at these levels. For instance, how realistic is the goal that by 2030 all the world's children will be proficient? Which categories of countries, or world regions, must be prioritised if global targets are to be met? How do demographic trends, in particular population growth, influence global proficiency statistics? Though the tool is not specifically designed for national planners, it should still be of interest to them insofar as its methods would to a large extent be applicable even for a national model.

A key simplifying assumption used in the tool is that in each country, proficiency is distributed normally. The justification for this is explained in section 4.

Section 5 discusses the merits and limitations of the dataset produced by Altinok for the UIS on country-level percentage proficient values.

Section 6 explains the rationale behind the work that has occurred previously on determining minimum proficiency levels, and what levels are implemented in the Excel tool. Specifically, a level corresponding to a score of 400 in TIMSS, plus a second much lower level of around 250 on the TIMSS scale, are explored. Given

clear problems with the latter, most of the projections and discussions focus on the higher minimum proficiency level.

Section 7 describes how Altinok's data were processed in order to prepare them for use in the Excel tool. As a starting point, patterns in Altinok's data are discussed. Why anomalous values for a couple of major countries had to be changed is explained. How values not available in Altinok's data were imputed is also described. This involved, firstly, imputations using just Altinok's data, for instance where countries had values for one education level, but not another. Secondly, it involved using external country-level per capita income to impute proficiency statistics. A further adjustment process involved bringing in UIS statistics on out-of-school children and assigning these children a non-proficient status. Examining and verifying the existing proficiency statistics reveals interesting opportunities for moving towards the SDG target. For instance, the relative under-performance of boys in the Northern Africa and Western Asia region should be addressed, as well as the relatively weak performance in mathematics of girls in the Latin America and Caribbean region.

Section 8 describes how the 2019 World Population Prospects demographic projections, which cover 2020 to 2100, were prepared for use in the tool.

Section 9 describes the structure of the Excel tool. The user does not need to specify many inputs to generate a scenario. A key input is the expected rate of progress, in terms of proportions of a standard deviation per year, relative to various points of departure. A related report, entitled How Fast Can Levels of Proficiency Improve, provides guidance on what rates of progress are plausible, given recent historical trends. It is assumed that less developed countries tend to see faster improvements than developed countries. Rates of progress can vary across each of the seven UNESCO world regions. The tool takes the baseline percentage proficient for each country, and constructs a normal curve that corresponds to this, and then weights this curve by the baseline population. For future years, the normal curve moves rightwards, in line with the expected improvement. Moreover, the normal curve expands or shrinks in line with the population projections. Nothing in the tool is hidden from the user, or locked. Key outputs of the tool include a set of graphs describing global and regional distributions of proficiency in the base year and in future years (the user selects the latter, and they may go as far as 2100).

Section 10 discusses margins of error. This is a complex matter given the multitude of factors that influence the confidence intervals around the regional and global proficiency statistics calculated for the report. As one might expect, confidence intervals are wider for developing countries, where testing systems are often not set at the level of the country, or where there is no test data and statistics must be imputed using other data. Confidence intervals must be taken into account when statistics are interpreted, yet they are not so wide that they invalidate the envisaged monitoring. To illustrate, the global statistic that 57% of lower primary children are proficient in the 2015 baseline carries a 95% confidence interval of about 54.5% to 59.5%.

Section 11 discusses findings emerging from the tool. A key finding is that differences across world regions in the demographic trends are large enough to influence substantially the aggregate proficiency situation. Specifically, high population growth in the region with the lowest proficiency statistics, namely sub-Saharan Africa, poses a special challenge. Much effort will need to go towards improving educational outcomes in this region if globally a movement towards the SDG on learning proficiency is to be maintained. In what is probably an unlikely scenario, where there is no progress in any country with respect to the percentage of children who are proficient, the global percentage of proficient children would move backwards, from 57% proficient in 2015 to 53% proficient in 2030, simply as a result of demographic shifts (in this example, lower primary reading is considered). In fact, a minimum degree of improvement at the country level, amounting to around 0.01 of a standard deviation a year across all countries, would be necessary simply to maintain the global figure of 57%. Improvements beyond 0.01 of a standard deviation would be required to allow the global figure to rise.

Contents

1 Introduction

This report is accompanied by another report, titled *How fast can levels of proficiency improve? Examining historical trends to inform SDG 4.1.1 scenarios*, and an Excel tool.

This work arose out of a need to bring about more certainty with respect to possible future trajectories and targets relating to Sustainable Development Goal (SDG) Indicator 4.1.1, which reads as follows[1]:

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

Given there are three school levels, two subjects and two sexes, one can think of there being a requirement for 18 different statistics. Six indicators are defined by school level and subject, and each of these six has three versions: for females, males and both sexes.

The actual goal seems to envisage attainment of a minimum level of proficiency for all the world's children and adolescents by 2030. That is clearly an aspirational target. What one could reasonably expect to achieve by 2030, and beyond that, is not clear. The aim of the current report is to provide greater clarity in this regard, using the available test and demographic data. The accompanying report looks back at what degree of improvement has occurred in the past[2]. The current report looks forward to 2030, and beyond.

While the immediate aim of this report is to clarify matters relating to indicator 4.1.1, a more indirect aim is to add to the limited but growing stock of tools and knowledge needed for effective education planning.

2 Learning from other areas: targets for enrolment, poverty and emissions

Projections into the future at the national and global levels in areas such as education, poverty and the environment are made for several reasons. Citizens, but also public and private investors, want to know what the future will look like if trends continue along historical paths. They also want to know how historical paths, or business-as-usual (BAU), could be deliberately changed to produce different futures. In part, making projections is about attempting to clarify key factors behind particular trends. Projections, which are the outcome of some policy to change the BAU trajectory, may be used for accountability purposes. Targets can be set based on the projections, which can then be used to assess to the degree to which individual countries, or the world as a whole, have succeeded in what they set out to do.

The future cannot be predicted perfectly. One obvious way of testing the reliability of a projection methodology is to apply it to an earlier period and then to see the degree to which it predicts what actually happened. For example, a method applied in 2015 to project 2015 to 2030 trends, could be applied retroactively to the 2000 to 2015 period, using only baseline information available to an analyst in 2000, to see whether predicted 2015 values were reasonably close to actual 2015 values. This type of testing is not common, in part because of the additional work it entails. A further explanation might be that people believe the initial projections are reliable enough. Or projections of an unknown level of reliability may be

[1] United Nations, 2017b.
[2] UIS, 2019.

accepted because the projections display what people would like to see, and serve the desired signalling purpose.

Projections and especially targets are highly political. There is often pressure for the producers of projections to insert overly optimistic assumptions into their models. This might be seen as necessary to create a sense of hope, and to pressurise governments and others to work especially hard at bringing about change.

Below, projections are discussed in three areas: enrolments, poverty and emissions. The aim here is to assess how lessons from these other areas could inform projections and target-setting in the area of learning outcomes, specifically proficiency as understood by indicator 4.1.1. There are two key questions. How have projections contributed to our understanding of what actions to take now? How were global and national targets set, how realistic were they, and how were they intended to encourage the desired policies and action?

2.1 Enrolments

Several projections were made around 2000 of attainment of the Millennium Development Goal (MDG) relating to Universal Primary Education (UPE) by 2015. The 2002 UNESCO Global Monitoring Report provides a summary of three sets of projections[3]. Projections helped to bring to the fore definitional issues. For instance, the World Bank's projections assumed a child had to complete six years of schooling to be considered to have received primary schooling. Yet many schooling systems in the world have fewer or more than six years of primary schooling, meaning the projections focussed on a global but somewhat theoretical notion of what constitutes primary schooling. Variations across countries in the structure of their schooling systems similarly complicate the discussion around indicator 4.1.1[4]. For example, should the level of proficiency 'at the end of primary' focus on the official last grade of primary schooling in each country, resulting in the aggregation of what are essentially different national indicators?

UPE focussed on participation in schooling, the implicit assumption being that this participation would lead to learning and a more education. To some extent, the projections acknowledged that this assumption may not hold, and that measures to improve the quality of schooling needed to be built into the projections[5]. The SDGs, on the other hand, focus explicitly on access to quality schooling. The concerns around quality expressed in relation to UPE were justified. While there appears to be no evidence that the quality of schooling deteriorated, as a result of UPE, for socio-economic categories of students who were already enjoying schooling in 2000, there is evidence that in many developing countries those communities which gained access to schooling through UPE often gained very little in terms of actual education[6].

The projections assumed that the target of universal primary education, meaning 100% completion of primary using the World Bank's definition, would be achieved by 2015. The unknown variable was the required budget, and the financial commitment required of donors. The assumption of successful attainment of the target was ambitious, but arguably realistic. In 2000, the adjusted net enrolment ratio (ANER) for the primary level, for the world as a whole, was around 83%. Put differently, around 17% of

[3] UNESCO, 2002: 134.
[4] UIS, 2018: 14.
[5] Mingat *et al*, 2002: 14.
[6] Taylor and Spaull, 2015; Birchler and Michaelowa, 2016.

children who should be attending primary school were not. In fact, as will be seen below, the global target was far more attainable than the 2030 *quality* target set for the primary level through SDG Target 4.1 in 2015. By 2015, the ANER had risen to 91%[7]. This improvement was achieved largely through a rapid decline in the number of out-of-school children in India, from around 50 to 3 million, an effect achieved mostly through relatively effective national policies[8]. In countries where development aid played an important role in promoting UPE, a focus on physical facilities, including the building of new schools, seems to have facilitated progress. The UPE projections included estimates of capital costs. The projections thus helped to reinforce an understanding of the practical steps needed[9].

2.2 Poverty

The United Nations Millennium Declaration of 2000 committed the world 'To halve, by the year 2015, the proportion of the world's people whose income is less than one dollar a day'[10]. At the time, rather basic and linear projections, which essentially projected forward the trend from 1990 to 2000, envisaged a reduction in the percentage of the world living on less than one dollar a head from around 29% in 1990 to 15% in 2015[11]. This would indeed represent a halving of the problem (reports from around 2000 make it clear that 1990, not 2000, was to be used as the baseline). Available reports from the time break the projections down by world region, and these breakdowns reveal that sub-Saharan Africa was expected to reduce poverty between 2000 and 2015 even faster than it had done during the 1990 to 2000 period. The 2015 United Nations report on MDG progress points to an actual reduction in the world from 36% to 12% between 1990 and 2015, with somewhat different measurements of poverty used[12]. It appears the world more than halved extreme poverty. However, there was one region that failed to halve poverty, and that was sub-Saharan Africa, where the extreme poverty rate dropped from 57% in 1990 to 41% in 2015.

More recent projections of poverty extending up to 2030 offer more interesting insights into the poverty reduction process. For instance, Lakner *et al* (2019) represent one of several attempts to gauge the combined impact of economic growth and income redistribution on poverty reduction. Impacts tend to rely heavily the shape of each country's initial income distribution. Reducing income inequality tends to accelerate poverty reduction, as one would expect, but this is not always the case.

Projections of poverty reduction have relevance for the current report insofar as improving the percentage of proficient children is about moving a distribution across a minimum point, a poverty line or minimum proficiency level. Proficiency, unlike enrolment, exists along a distribution. It is not a simple binary concept. The non-proficient may be almost proficient, or may be very far from being proficient. However, the distribution of income is in many ways different to the distribution of cognitive skills. The latter tends to be distributed normally, while income is not (see discussion in section 4 below). Moreover, cognitive skills cannot be redistributed towards the disadvantaged, in the way income can. Yet the parallels between poverty and level of skills acquired serve as a reminder that policy action specifically aimed at reducing educational inequality is important. Moreover, in both areas, poverty and skills, there is an important tension between, on the one hand, attempting to cover a wide range of issues through many indicators

---

[7] UNESCO, 2002: 359; UNESCO, 2017: 320; United Nations, 2015: 24.
[8] UNESCO, 2015: 76.
[9] Birchler and Michaelowa, 2016; UNESCO, 2002: 136.
[10] United Nations, 2000.
[11] IMF, 2000: 5-6.
[12] United Nations, 2015: 14.

and, on the other, reducing the complexity of the monitoring process. To illustrate, the SDG on poverty reduction places a strong emphasis on the various dimensions of poverty, including access to basic services and access to a 'safety net' in the form of a publicly funded social protection system. While the SDGs on learning proficiency focus just on reading and mathematics, there is pressure to broaden the scope to include other school subjects, and even areas of life skills such as financial literacy. PISA[13] began focussing on the latter in 2012[14].

2.3 Emissions

Projections on emissions reductions, needed to slow down climate change, are today probably supported by more analysis and documentation than any other projections in existence. Climate change is such an all-encompassing phenomenon that all policy areas have begun to take into account its effect. UNESCO (2016) has developed guidelines on how schooling systems can become 'climate-ready'.

From an SDG 4.1.1 perspective, what is interesting about the Paris Agreement on climate change are the rather detailed requirements relating to monitoring and the setting of targets. Countries set their emissions reduction targets in line with national projections and intended policy changes, in a manner that contributes to the global target of less than two degrees of warming. However, there is no system ensuring that all national targets together realise the global target. The emphasis is thus strongly on countries doing the best they possibly can. The ability to track historical trends underpins the ability to set targets. In recognition of differing levels of technical capacity across countries, the internationally accepted guidelines permit three different methodologies, from most to least demanding in the terms of national capacity, and from most to least precise[15]. National reports on emissions are submitted to the UNFCCC[16] as often as possible. These reports must provide information in standard tables, but also a supporting narrative. In the case of developing countries, instead of stipulating a fixed frequency for reporting, the emphasis is instead on building capacity for accurate reporting. Reports are reviewed by the UNFCCC. For instance, a 2019 36-page review was prepared to assess Kazakhstan's 2018 emissions report. The review[17], conducted by five experts from different countries, who spent five days in the country, focussed on the transparency of the methods used, comparability of statistics over time, and to some extent the effectiveness of national policies aimed at reducing emissions. Kazakhstan was given the opportunity to comment on the review before it was published online.

Two key global repositories of national emissions data exist. On the one hand, the UNFCCC releases the data reported by countries. Secondly, the World Resources Institute (WRI) maintains the online Climate Analysis Indicators Tool (CAIT), which draws extensively from UNFCCC, but also other sources, with a view to producing national statistics that are as accurate as possible. UNFCCC and CAIT values tend to differ, with the latter being considered more reliable[18], and hence being used most extensively by researchers.

---

[13] Programme for International Student Assessment.
[14] OECD, 2017.
[15] Intergovernmental Panel on Climate Change, 2006: 8.
[16] United Nations Framework Convention on Climate Change.
[17] Report titled *Report on the technical review of the seventh national communication of Kazakhstan*, dated 26 July 2019, and available at https://unfccc.int/sites/default/files/resource/idr7_KAZ.pdf (accessed August 2019).
[18] 2016 report titled *Global greenhouse gas emissions* by the US Environmental Protection Agency, available at https://www.epa.gov/sites/production/files/2016-08/documents/global-ghg-emissions_documentation.pdf (accessed August 2019).

Historical CAIT values are more likely to be revised than historical UNFCCC values as new and better information emerges. A WRI online facility allows for easy comparison of UNFCCC against CAIT trends[19].

Target-setting and monitoring in the area of learning outcomes is complex. It is less complex than in the area of emissions, but learning outcomes imply far greater levels of complexity than traditional education indicators such as enrolment, grade repetition or pupil-teacher ratios. An emphasis on statistical capacity building is clearly needed. An online repository of reports submitted by countries, to underpin published statistics, as well as a system for reviewing submitted reports, could be a way of ensuring that capacity is built.

2.4 Lessons from the comparison

The trajectory for proficiency is arguably more difficult to model than that for any of the three other indicators discussed above, given the current data and knowledge about the topic. Firstly, knowledge about country-specific and global trends in the past is still very limited, in part because the existing data cover only a part of the globe, and probably that part most likely to see progress (see section 7.1 below). Secondly, the factors influencing change, in particular at the level of whole countries, as opposed to groups of schools subjected to 'experiments', are poorly understood. All this limits what can be said about the likely future of proficiency. Yet there is scope for a far better knowledge base to be created. Distributions of learning outcomes are not inherently more difficult to measure than distributions of income. Monitoring systems, for example regional programmes such as SACMEQ[20] and PASEC[21], need to be strengthened, and more countries need to participate in them. The work of which the current report is a part takes the knowledge base forward, yet there is much more to be done. While the projections presented below are in certain respects crude, the thinking that informs target-setting currently within education systems is considerably cruder, and produces targets that tend not to be helpful[22].

Going forward, there are probably opportunities for causal factors to be brought into projections of learning proficiency in a formal manner, in the way causal factors have increasingly been built into projections of poverty. Such work should be informed by the limited work drawing from international datasets, which examines systemic factors associated with higher test score averages[23]. A key limitation of this work is that, given its cross-sectional data focus, it fails to properly take into account historical and institutional factors. Investigations into countries that have progressed rapidly in the international testing programmes, aimed at understanding the improvement, can help to fill this gap. An example of such an investigation would be Bruns *et al* (2012), who examine factors behind Brazil's educational improvements.

With respect to building capacity at the national level, UNFCCC processes in relation to emissions point to interesting opportunities. These processes explicitly acknowledge that countries have vastly different levels of capacity to report, and hence outline three different routes countries can follow, depending on their capacity. UNFCCC reviews of the reports submitted by countries could be adapted and replicated for the complex area of learning outcomes, in particular where countries are developing national assessment

---

[19] www.climatewatchdata.org.
[20] Southern and Eastern Africa Consortium for Monitoring Educational Quality.
[21] Le Programme d'analyse des systèmes éducatifs de la CONFEMEN (Programme for the Analysis of Educational Systems).
[22] See UIS (2019).
[23] For instance Wößmann (2005) and Hanushek and Woessman (2007).

systems to complement the international ones. Lastly, having two levels of reporting at the global level, one with 'raw' statistics submitted officially by countries, and another where these statistics have been adjusted by experts, seems worth considering in the area of learning proficiency statistics.

3 What the Excel tool does and does not do

A key purpose of the Excel tool, and the preparatory data work described in this report, is to combine country-level data on three things – proficiency levels, the out-of-school problem and population growth – and a number of global or regional assumptions about the way forward, to produce baseline and future statistics relating to the percentage of children who are proficient in reading or mathematics, with breakdowns as specified in indicator 4.1.1. The focus is on producing statistics at the global level, and at the level of the seven UNESCO world regions, used in, for instance, the 2019 Global Education Monitoring Report[24].

More broadly, the tool is intended to facilitate the answering of important questions. If the goal that by 2030 all the world's children will be proficient is over-optimistic and aspirational, how over-optimistic is it? When might this goal be achieved beyond 2030? Which categories of countries, or world regions, must be prioritised if global targets are to be met? What do gender differences say about how overall improvement should be organised? How do demographic trends, in particular population growth, influence global proficiency statistics?

On methodology, the work underpinning the tool, and explained in the current report and the additional report accompanying this one, is aimed at answering certain questions. What is the shape of the distribution of test scores, and how does this influence things such as the possible speed of improvement? How should progress be measured when data come from a variety of monitoring systems?

One thing the tool does *not* do is describe with much accuracy the baseline situations in individual countries. The global dataset compiled for the projections is the result of adjustments that could skew the actual baseline in individual countries, as well as imputations for, at each of the three school levels, between one- and two-thirds of the world's child age cohort. By extension, the tool does not set out to recommend to individual countries what their SDG 4.1.1 targets should be. Official guides on the SDGs make it clear that countries should set their own targets[25]. However, general patterns produced by the tool, apart from providing meaningful values for the globe and world regions, should also be informative for national planners. In particular, they provide a basis for more evidence-based planning in an area where targets are often set politically and aspirationally. While aspirational targets can inspire hope, they can also produce an unjustified sense of failure when they are not met. Understanding the patterns discussed in this report is thus important for analysts wishing to detect progress, where progress is often missed.

Despite the emphasis in the accompanying report on taking into account the effect of grade repetition, and changes in this, on proficiency trends, this is not built into the Excel tool. Grade repetition effects were

---

[24] UNESCO, 2019.
[25] United Nations, 2017a.

avoided to limit complexity in the model. Moreover, although clearly repetition has an effect, it is very unlikely to alter the overall findings of the work presented below.

The tool could form the basis for a different tool, designed for use by national planners. Such a tool should be aimed at assisting a national authority to understand past trends in an assessment system, possibly a national assessment, as well as the distribution of student performance. Properly structured knowledge on this would facilitate national target-setting in much the way the current Excel tool does this at a global level.

4 Assuming a normal distribution of proficiency

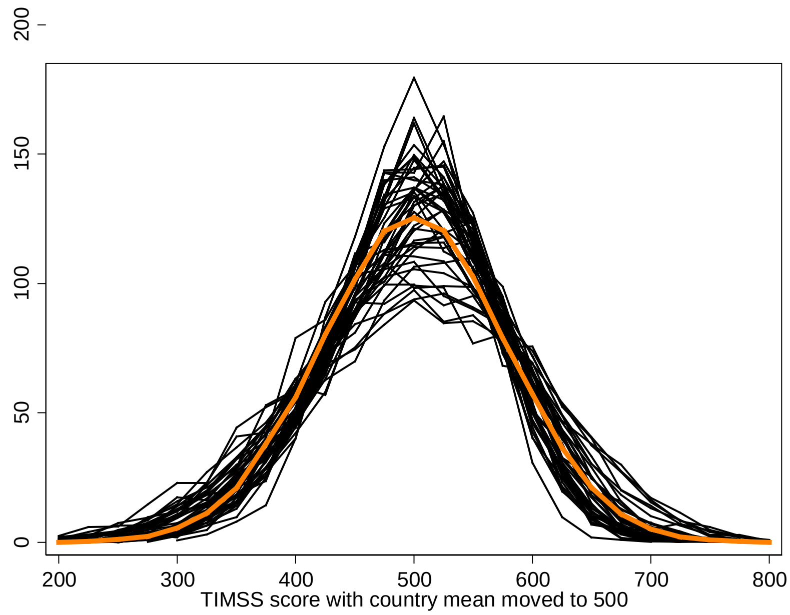
This section argues that scores in existing testing systems display a sufficiently normal distribution for it to be assumed within the projection model that change over time occurs in terms of a normal distribution of learning proficiencies which moves to the right, or towards better levels of learning.

The literature warns against assuming normality in test scores too easily, specifically where this assumption is used in regression-based test score analyses[26]. The projection model described below uses the assumption of normality for a very simple purpose: to gauge changes by year in the number of children who pass a minimum proficiency level as a country's test score mean improves. To illustrate, if the distribution of scores is uniform, meaning for instance that as many children in a baseline year have a score of 420 as 425, then the distribution would be flat and constant improvements in the mean would result in the same number of children crossing the proficiency threshold each year. However, we know that the distribution is not uniform, or flat. But then what do the gradients of the curve look like, and how normal is the distribution?

Figure 1 illustrates the distribution of mathematics scores in TIMSS27 Grade 4 of 2015. The shape of each country's original curve is retained, but the curve has been moved rightwards or leftwards to produce a mean of 500 as the aim here is to compare just the height and shape of the curves. Each black curve represents one of 41 countries, and weights have been adjusted to produce the same number of students per country. The orange curve is a normal curve with a mean of 500 and a standard deviation of 79 TIMSS points. If the standard deviation is calculated for each of the 41 countries, the mean across these 41 statistics is 79. The orange curve is thus what one could expect all countries to display if they all had normally distributed test scores and the same standard deviation.

[26] Ho and Yu, 2015.
[27] Trends in International Mathematics and Science Study.

Figure 1: Normality of 2015 TIMSS Grade 4 mathematics scores

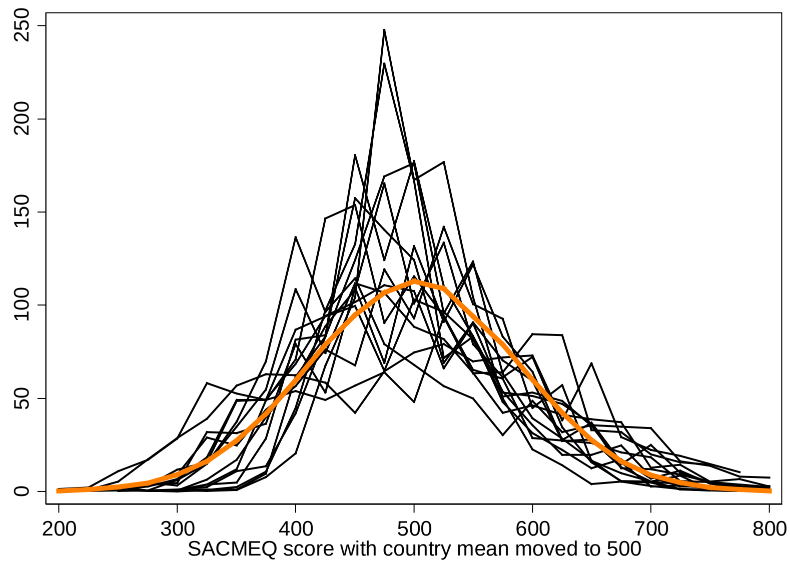


Source: TIMSS microdata available at https://timssandpirls.bc.edu. The first of the five plausible values was used.
Note: Vertical axis is weighted students at z score intervals of 0.25 with the weight multiplied by a factor to produce 1000 students per country.

This visual inspection suggests curves are remarkably normal. One should remember that even country-specific curves with peaks considerably higher or lower than the peak of the orange curve could still be normal. Further analysis indicates that no countries pass a strict normality test, specifically a skewness and kurtosis test for normality ('sktest' in Stata). However, skewness and kurtosis statistics per country viewed separately confirm that the curves are close to being normal. Skewness for the 41 countries ranges from -0.45 to 0.17, the mean across countries being -0.19. A perfectly normal curve would produce a skewness of zero. State-level testing programmes in the United States produce measures of skewness which cluster around minus 0.5 to positive 0.5 (Ho and Yu, 2015: 371). The TIMSS countries fall within this range, meaning there are no individual countries displaying exceptionally non-normal asymmetries. The mean of -0.19 points to curves leaning slightly to the right, which is visible in the graph. The kurtosis statistics for the 41 countries, which fall within the range 2.6 to 3.5 (mean of 3.0) are also not unusual for test data. A kurtosis statistic of 3.0 would emerge from a perfectly normal curve. Importantly, the skewness and kurtosis statistics do not vary noticeably depending on whether one is looking at high income OECD countries or not.

SACMEQ 2007 reading test score data were analysed in the same manner to produce **Figure 2**. The 15 countries here produce skewness statistics ranging from -0.23 to 1.10, with a mean of 0.41, and kurtosis measures ranging from 2.3 to 5.5, with a mean of 3.4. The ranges for both statistics are larger than those found using TIMSS. Moreover, both means deviate from the normal curve values to a greater degree in the case of SACMEQ. The positive mean skewness of 0.41 points to leftward leaning curves. This can be seen in Figure 2 if one compares the country curves to the orange normal curve. The lesser normality of the SACMEQ curves are likely to be related to measurement anomalies which have been observed in the SACMEQ data[28], and which are to be expected, given the lower level of resourcing and shorter history of SACMEQ, relative to TIMSS.

Figure 2: Normality of 2007 SACMEQ Grade 6 reading scores

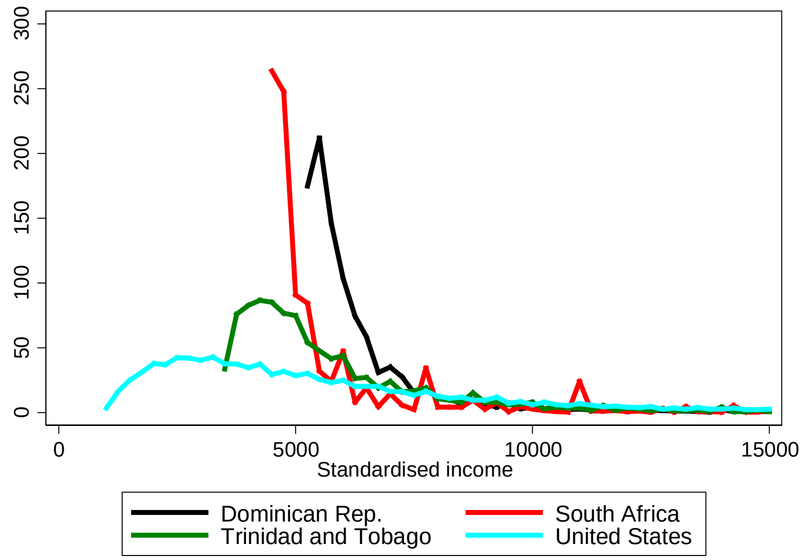


Source: SACMEQ III microdata.

The distribution of learning outcomes is clearly more straightforward than the distribution of income. Figure **3** illustrates the distribution of per capita income in four countries, where each country's statistics assume the same mean and standard deviation, to allow for just the shape of the distribution to be compared. Some countries, such as South Africa, do not have a clearly downward sloping 'tail' on the left of the distribution, while other countries, such as Trinidad and Tobago, do. Thus, analysts projecting levels of poverty are often concerned about the 'depth' of poverty, or the shape of the distribution among those below the poverty line[29]. Given the relative predictability of the distribution of learning outcomes, such concerns are far less important in education.

[28] Crouch and Gustafsson, 2018.
[29] Lakner *et al*, 2019: 16.

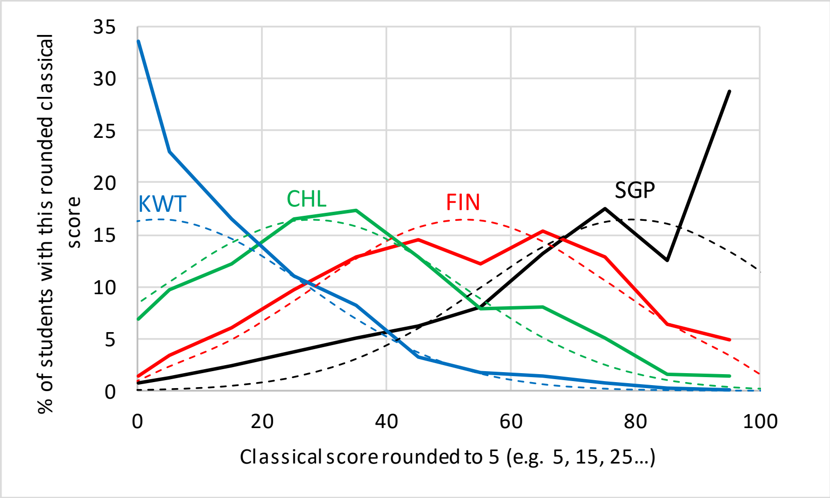
Figure 3: Income distributions



Source: IPUMS, Minnesota Population Center, accessed 2019. Providers of underlying data: National Statistics Office, Dominican Republic; Statistics South Africa; Central Statistical Office, Trinidad and Tobago; Bureau of the Census, United States.
Note: For each country, per capita income figures were standardised to produce a mean of 7000 and a standard deviation of 7000.

Importantly, while the item response theory (IRT) scores provided by testing systems such as TIMSS and SACMEQ display a largely normal distribution, this may be less true of the underlying classical scores. In particular, where floor and ceiling effects prevail, the distribution of classical scores may be obviously non-normal. This can be seen in **Figure 4** below. Here Kuwait (KWT), a poor performer in TIMSS, displays a classical test score distribution with a peak near zero, while Singapore displays the opposite, namely a peak near 100% correct, or the ceiling. One of the effects of IRT scoring is to bring about more differentiation between students near the floor and ceiling, thus producing distributions which are more normal (illustrated by the dotted lines in the graph).

Figure 4: Floor and ceiling adjustments in TIMSS 2015 Grade 4 mathematics



Source: From Gustafsson, 2019.

The smoothing brought about by IRT scoring is often not well understood. It is not as if this scoring creates differentiation among students near, say, the floor when none actually exists. The differentiation is based on which questions students provided correct responses to. This is not to say that serious floor effects do not sometimes confound the analysis. Tests whose difficulty level is clearly too high for a specific country will produce floor effects which will confound, for instance, proficiency statistics. The little research that has looked into floor effects suggests that it is in fact in assessment programmes designed for developing countries, specifically LLECE[30] and SACMEQ, where floor effects do emerge as concerning, although these programmes were in part set up to deal with such effects. However, while floor effects do sometimes distort proficiency statistics, they are not large enough to distort, say, the ranking of countries substantially[31].

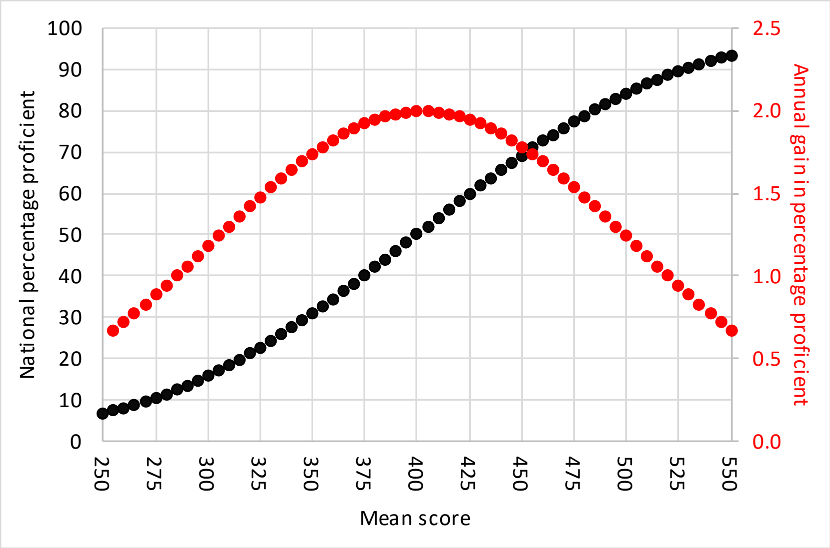
How should out-of-school children be taken into account? In the projection model presented in subsequent sections, it is assumed that these children occupy mostly the left-hand end of the normal curve. It is often argued that out-of-school children should be assumed to display zero proficiency – see for instance UIS (2017b). A more nuanced, and statistically simpler, assumption is employed in the current report. It is assumed, firstly, that there are proficiency differences among the out-of-school, in part because poverty in the form of malnutrition would have affected children differently, giving some children more cognitive and learning opportunities than others. Secondly, it is assumed that even those children in school could display levels of proficiency that are at least as low as those found among the out-of-school, due to poor quality schooling. It is thus assumed that the normal curve accommodates all children, whether they are in school or not. What is therefore avoided is a normal curve for in-school children, with a separate peak, at what one could consider a zero-proficiency level, for out-of-school children.

---

[30] Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education).
[31] Gustafsson, 2019.

A normal distribution of proficiency implies that one can expect faster improvements to the percentage proficient indicator in initial years, in the case of a country starting from a low base, and slower improvements after the peak of the normal curve has passed the minimum proficiency threshold. This is illustrated in **Figure 5**, which is constructed using data from the projection tool. The mean score of the illustrative country improves from a score of 250 to 550 during the 2015 to 2100 period. The minimum proficiency level is 400. Up to the point at which the mean reaches 400, the country experiences gains in the indicator that are larger for each year. After this point, the annual gains start shrinking.

Figure 5: Rate of global change up to 2100



Note: The red curve should be read against the right-hand vertical axis.

To conclude this section, the test microdata suggest that test scores, at least the IRT scores used in international comparisons, are distributed to a sufficiently normal degree to permit a projection model, such as the one presented below, which is based on a series of country-specific normal curves.

5 The merits and limitations of Altinok's harmonised dataset

The harmonised dataset produced by Altinok and used for the current work was described in the accompanying report. The dataset in question is available on the UIS website[32], with details on its construction appearing in Altinok (2017) and UIS (2018). It was produced to provide a set of baseline statistics for monitoring indicator 4.1.1. The dataset is undoubtedly useful insofar as it produces about as much comparability across different testing programmes, and hence across countries, as is possible given what data are currently available. As pointed out by Altinok, there are serious limitations, and having a worldwide testing programme, or even some new 'bridging' tests to improve the comparability of results across programmes, would be ideal.

Key limitations, and their implications for the current work, are described below.

[32] http://tcg.uis.unesco.org/data-resources. Accessed June 2019.

A limitation Altinok (2017: 23) focusses on is possible problems in the process whereby testing programmes are rescaled to fit one single scale, namely the TIMSS scale. There are several reasons why one country participating in two testing programmes, where the two student distributions are joined to produce a common scale, would not result in a score of *x* in one assessment that means the same as *x* in the other assessment. If the two samples do not represent the same population, for instance if they represent different grades, *x* clearly would not mean exactly the same across the two in curriculum terms, though *x* might mean something similar in terms of student ranking. If the two assessments test different aspects of the same subject, *x* in one assessment will not mean *x* in the other. Lastly, if the two assessments occurred in different years, and progress has been occurring, one should not expect the two distributions to equal each other in terms of the quality of learning outcomes.

A further limitation is that Altinok's dataset did not take into account anomalies within assessment programmes which have implications for comparability *within* programmes. For example, the presence of different metrics *within* the PIRLS[33] programme in the case of a couple of countries was not taken into account. Moreover, within programmes a country participating in a different grade results in comparability problems. Considering regions within a country as representative of the country is also problematic. These issues are addressed in section 7.2 below.

One matter which is unclear in Altinok (2017), is why in the case of the Latin American LLECE programme, for 2006 and 2013 only Grade 6 testing was considered, and not also the LLECE Grade 3 testing, given that opportunities existed for 'doublooning', as Chile participated in both LLECE and TIMSS Grade 4 testing.

6 Choosing a global minimum proficiency level

The accompanying report discussed how minimum proficiency levels are derived *within* assessment programmes. Here the matter of a single *global* level is discussed.

Indicator 4.1.1 requires selecting a specific minimum proficiency level, which inevitably involves some subjectivity around what should be defined as proficient at the global level. For the purposes of the current study, Altinok's approach was followed. Altinok (2017: 42, 55) considered the TIMSS 'low international benchmark' of 400 TIMSS points as *the* global minimum proficiency level, MPL. By extension, a score of 400 in the IEA's[34] other testing programme, PIRLS, was also considered an MPL.

Altinok uses the TIMSS-based MPL of 400, regardless of whether the TIMSS testing is occurring at the Grade 4 or Grade 8 level. By design, 400 in TIMSS represents roughly a specific percentile in the overall test-taker distribution, or at least the distribution as it existed in an original wave of TIMSS. Thus using the same 400 MPL at both the Grades 4 and 8 levels means a percentile-based, rather than a truly competency-based, criterion is used, within TIMSS, but by extension also by Altinok. Put differently, it is assumed that if a country's percentage proficient statistic at Grade 4 is *x*, it will be around *x* in Grade 8 too. There is thus no assumption that students increasingly fall behind as they move up the grades. The fact that despite this, patterns in the Altinok data *do* often point to lower levels of proficiency at a higher level of schooling system, is discussed in section 7.3.

[33] Progress in International Reading Literacy Study.
[34] International Association for the Evaluation of Educational Achievement.

In addition, Altinok proposed using a lower proficiency level, the 'SACMEQ benchmark', referred to below as the basic proficiency level, or BPL, in the case of developing countries, so that the gap between the baseline and the attainment of proficiency for all would not appear extremely large. This benchmark lies at a score of around 270, using the TIMSS metric (Altinok, 2017: 55).

7 A baseline drawing largely from Altinok's harmonised data

This section describes in considerable depth how lower primary reading statistics were prepared for the tool. This level (lower primary) and domain (reading) are obviously of particular importance as this is the foundation for all learning in schools. Statistics for other levels and domains are described in less depth as to a large extent their calculation draws from the methods developed for lower primary reading.

7.1 The global coverage of Altinok's data

For the purposes of the current analysis, 227 'countries' are assumed. These include 226 entities in the column 'Country' in the UIS online tables dealing with the SDG indicators, as these tables stood in June 2019[35]. Among the 226 are Palestine, 'China, Hong Kong Special Administrative Region' and 'China, Macao Special Administrative Region'. Chinese Taipei, which participates in PISA and TIMSS was then added. The resulting set would cover virtually the entire human population, which was a prerequisite for the generation of scenarios.

Population values per country used for the following tables are the most recent ones available in UIS.Stat in June 2019 – the values for most countries were for 2017. The population aged 0 to 14 was used, as this is roughly the population we are concerned with here. For 30 countries without UIS.data population values, mainly very small countries, population was estimated using alternative sources, for instance the United Nations Population Division (UNPD).

Table 1: Full set of 'countries'

UNESCO 2019 region	Countries	Population aged 0 to 14 (thous.)
Sub-Saharan Africa	50	436,846
Northern Africa and Western Asia	23	154,071
Central and Southern Asia	14	557,134
Eastern and South-eastern Asia	19	453,059
Oceania	23	9,540
Latin America and the Caribbean	48	161,169
Europe and Northern America	50	191,776
Total	227	1,963,594

Table 1 above reflects the universe of countries and populations used for the denominator in the two tables which follow. The regions are those of the 2019 Global Education Monitoring Report. **Table 2** and **Table 3**

[35] 226 was the result of the removal of 'Sudan (pre-secession)' – 'Sudan' and 'South Sudan' existed separately in the lists – and of the Holy See (which was assumed not to have children).

make it clear that mathematics is somewhat better covered than reading in the Altinok data: 79% of the global population for the former, but 73% for the latter[36]. End of primary is only covered in two regions, reflecting the presence of SACMEQ, PASEC and LLECE. Coverage of the critically important primary level, even for mathematics (42% of the population, counting both primary levels), is much lower than one would wish. This is partly due to the absence of data on China and India at this level, who jointly account for 32% of the world's child population. Yet 26% of the world's children are 'missing' in the test data and are outside of China and India. The secondary level population is far better covered than the primary level population. This reflects the widespread use of TIMSS and PISA at the secondary level.

Table 2: Coverage for reading (percentages)

	Lower prim.	End prim.	Any prim.	Lower sec.	Any level
TOTAL FOR COUNTRIES	24	22	44	34	54
Sub-Saharan Africa	4	60	60	2	60
Northern Africa and Western Asia	43	0	43	43	61
Central and Southern Asia	7	0	7	21	29
Eastern and South-eastern Asia	16	0	16	53	53
Oceania	9	0	9	9	9
Latin America and the Caribbean	10	40	44	27	46
Europe and Northern America	64	0	64	78	80
TOTAL FOR POPULATIONS	18	21	36	57	73
Sub-Saharan Africa	4	57	57	0	57
Northern Africa and Western Asia	31	0	31	30	43
Central and Southern Asia	3	0	3	68	71
Eastern and South-eastern Asia	16	0	16	86	86
Oceania	58	0	58	58	58
Latin America and the Caribbean	16	96	96	81	97
Europe and Northern America	83	0	83	92	92

One thing is noteworthy. Though Altinok's data consists of three tables, for the three school levels of the SDG indicator, Altinok's (2017) report is concerned with reporting on levels of proficiency of the primary level in general. This probably explains why the data table for the lower primary level does not make use of either the LLECE or PASEC source, though both these sources include both lower and upper primary testing. Altinok's data *does* draw from LLECE and PASEC at the upper primary level. At the lower primary level, the only source is PIRLS, for reading, and TIMSS, for mathematics. Had Altinok also drawn from, say, LLECE at

[36] In calculating these types of percentages, each country was weighted by the 0 to 14 population. Insofar as age cohorts do not differ much across ages within each country, this method would not unduly distort the global statistics, relative to a method where population totals for separate age cohorts were used.

the lower primary level, the percentage of countries covered in the Latin America and Caribbean region would have been considerably above the 10% seen in Table 2.

Table 3: Coverage for mathematics (percentages)

	Lower prim.	End prim.	Any prim.	Lower sec.	Any level
TOTAL FOR COUNTRIES	29	22	48	43	58
Sub-Saharan Africa	4	60	60	8	62
Northern Africa and Western Asia	70	0	70	83	87
Central and Southern Asia	14	0	14	29	29
Eastern and South-eastern Asia	42	0	42	63	63
Oceania	9	0	9	9	9
Latin America and the Caribbean	8	40	40	31	44
Europe and Northern America	64	0	64	84	84
TOTAL FOR POPULATIONS	23	21	42	66	79
Sub-Saharan Africa	4	57	57	7	60
Northern Africa and Western Asia	50	0	50	71	78
Central and Southern Asia	4	0	4	71	71
Eastern and South-eastern Asia	32	0	32	93	93
Oceania	58	0	58	58	58
Latin America and the Caribbean	12	96	96	84	97
Europe and Northern America	90	0	90	99	99

Table 4 permits a view of the time periods covered by the data. 1995 to 2016 is the overall range. For most regions, across the three levels, relatively recent data, 2015 or 2016, for at least some countries, are available.

Table 4: Year ranges for the Altinok data

	Min. year	Median year	Max. year
LOWER PRIMARY			
Sub-Saharan Africa	2011	2011	2016
Northern Africa and Western Asia	1995	2011	2016
Central and Southern Asia	1995	2007	2016
Eastern and South-eastern Asia	1995	2007	2016
Oceania	1995	2007	2016
Latin America and the Caribbean	2001	2009	2016
Europe and Northern America	1995	2007	2016
END OF PRIMARY			
Sub-Saharan Africa	1995	2006	2011
Latin America and the Caribbean	1997	2006	2013
LOWER SECONDARY			
Sub-Saharan Africa	1995	2008	2016
Northern Africa and Western Asia	1995	2009	2016
Central and Southern Asia	1995	2009	2016
Eastern and South-eastern Asia	1995	2009	2015
Oceania	1995	2006	2015
Latin America and the Caribbean	1995	2009	2015
Europe and Northern America	1995	2007	2015

7.2 Adjustment of certain values in Altinok's data

There appear to be two types of adjustment required for the Altinok proficiency statistics. These are explained below. Importantly, only the most recent Altinok value for each country, domain (subject) and level was used to establish the baseline described in subsequent sections. Hence, the two adjustments focus on these more recent values. Whether adjustments were required for non-recent values was not considered, as this would not impact on the construction of the baseline.

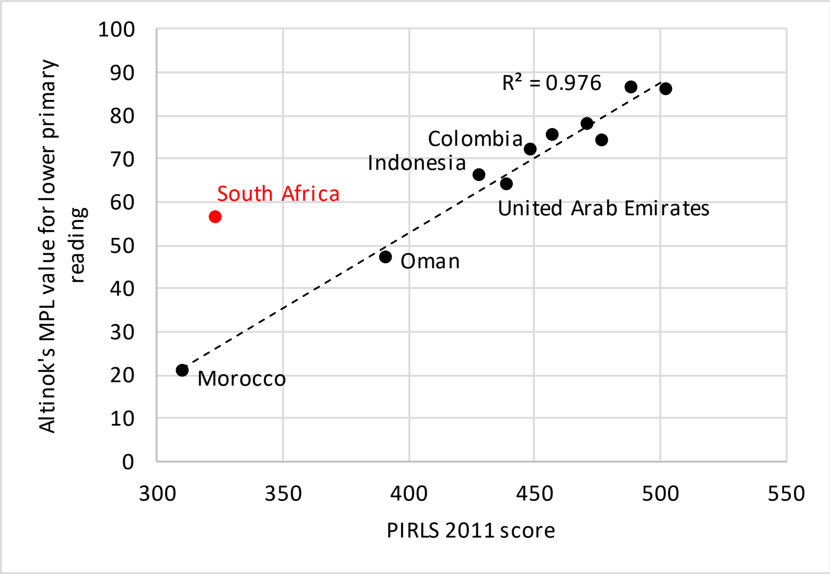
In the remainder of the current report, terms such as 'original Altinok data' are used to refer to the Altinok data *after* the adjustments described in this section were effected.

7.2.1 PIRLS-derived South Africa and Botswana values

The first adjustment applies to the 2011 lower primary reading values for two countries, South Africa and Botswana. These values draw from the PIRLS 2011 data. The problem here is that South Africa and Botswana did not participate in the regular PIRLS, but in the easier prePIRLS, for which the results were originally not comparable to regular PIRLS results. It is not clear how Altinok dealt with this lack of comparability, but what is clear is that using the South Africa and Botswana proficiency statistics as they appear in the Altinok dataset would be problematic and would produce inconsistent patterns for not just these two countries, but even other countries where imputations were influenced by South Africa and Botswana. As will be seen in section 7.4, these two countries are indeed particularly important in the imputation process.

That Altinok's South Africa and Botswana proficiency statistics are problematic is illustrated in **Figure 6** below. Here South Africa's PIRLS score, plotted against the horizontal axis, is not the prePIRLS score shown in the official PIRLS 2011 report, a score which is not comparable to the standard PIRLS scores used for most countries, but the original prePIRLS score (461 according to the official report[37]) converted to the standard PIRLS scale by the IEA. How this conversion occurred appears not to be documented in any publicly available report, but it is clear that the conversion occurred and that South Africa's score after this conversion was 323. It is this 323 which is used in Figure 6 for South Africa. This value appears in the global report for PIRLS 2016, specifically in a comparison to South Africa's 2016 PIRLS score of 320[38]. In 2016, South Africa also participated in an easier PIRLS test, called PIRLS Literacy, but the PIRLS Literacy score *was* expressed in terms of the general PIRLS metric, and only in that metric. It is not clear what Botswana's converted 2011 PIRLS score would be, as Botswana did not participate in PIRLS 2016. However, Botswana's performance in PIRLS 2011 was close to that of South Africa, with the country achieving a prePIRLS score of 463 (against South Africa's 461).

Figure 6: Problem with the prePIRLS alignment



Source: PIRLS scores represented by the horizontal axis (except for South Africa) are from Mullis et al (2012: 38).

The decision was taken to revise South Africa's percentage proficient statistic (using the minimum proficiency level, or MPL) for lower primary reading from the 57% seen in Altinok's data, to 26%, which would be in line with the trendline in Figure 6 driven by the data from regular PIRLS participants. Botswana's figure was similarly adjusted from 56% to 27%. Adjusting South Africa and Botswana's scores as described here would place these countries' lower primary reading values more in line with other values seen in Altinok's data, for

[37] Mullis *et al*, 2012: 39.
[38] Mullis *et al*, 2017: 33.

instance 22% and 25% for reading using the MPL for the same two countries at the *upper* primary level. For revised basic proficiency level (BPL) percentages, for lower primary reading, the MPL-BPL function described in section 7.4 was used.

Honduras is also a problem. This country displays a peculiarly high mean score in PIRLS 2011, and consequently carries a reading MPL percentage, for lower primary, in the Altinok data which is high, and at odds with Honduras's relative performance in LLECE. In the case of Honduras, this is because the country entered Grade 6 students, instead of Grade 4 students, in the regular PIRLS testing of 2011. Honduras's values in Altinok were not revised. Honduras is a small country which barely affects the global and regional patterns we are interested in, and moreover Honduras does not play a critical role in the imputation processes described in section 7.4.

7.2.2 The sub-national PISA sources for India and China

In 2015, the Chinese provinces of Jiangsu and Guangdong, and the municipalities of Beijing and Shanghai, participated in PISA. These four represent around 17% of the national population. In India in 2009, the states of Tamil Nadu and Himachal Pradesh participated in PISA. Obviously, the key question is how different the results would have been for China and India had the entire countries participated. This is of particular importance for the current analysis given the population sizes of China and India.

China's lower secondary MPL values in Altinok's dataset are 80% for reading and 88% for mathematics for 2015. A 2017 presentation by China's National Assessment Center of Education Quality, available online through the UIS, refers to an 'international linking trial' and states that this trial indicates that 'Mainland China ranked about 10th in the 65 participating countries in PISA 2012'[39]. This would place China on a par with Estonia in mathematics, or a 93% proficiency level using Altinok's lower secondary figures, and also on a par with Estonia in reading, or a 91% level. Estonia emerges in position ten in the official 2012 PIRLS report. Should China's reading and mathematics values be adjusted upward to match those of Estonia? The problem with this is that the four 2015 participants Jiangsu, Guangdong, Beijing and Shanghai display GDP per capita figures that are clearly above the average for China. It seems unlikely that these four would perform *worse* educationally than the rest of the country. In the absence of additional information, for instance on the international linking trial, it was decided to leave the China values in the Altinok dataset unchanged.

India's lower secondary MPL values in Altinok's dataset are 19% for reading and 25% for mathematics, the source year being 2009. Of the two Indian states participating, Tamil Nadu and Himachal Pradesh, Himachal Pradesh emerges as close to the Indian average according to a report drawing from India's own National Achievement Survey (NAS), run apparently for the first time in 2017[40]. Specifically, Grade 3 reading results from this survey produce a mean of 257 for India, 256 for Himachal Pradesh, a higher 274 for Tamil Nadu, with state mean scores ranging from 226 for Chhattisgarh to 281 for Tripura. A report covering non-OECD participants in PISA 2009 indicates that Himachal Pradesh and Kyrgyzstan were together the lowest performers in lower secondary reading in 2009[41]. Kyrgyzstan's lower secondary MPL values in Altinok's dataset are 18.6% for reading and 20.5% for mathematics, the source year being 2009. India's two values were adjusted downward to match those of Kyrgyzstan, not a large adjustment in the greater context.

---

[39] China: Ministry of Education, 2017.
[40] India: National Council of Education Research and Training, 2018.
[41] Walker, 2011: xiv.

Clearly, more work could be done with the existing microdata to arrive at a more informed adjustment. However, at this point nothing seems to suggest that a more informed adjustment would produce substantially different results.

7.3 School level and subject differences

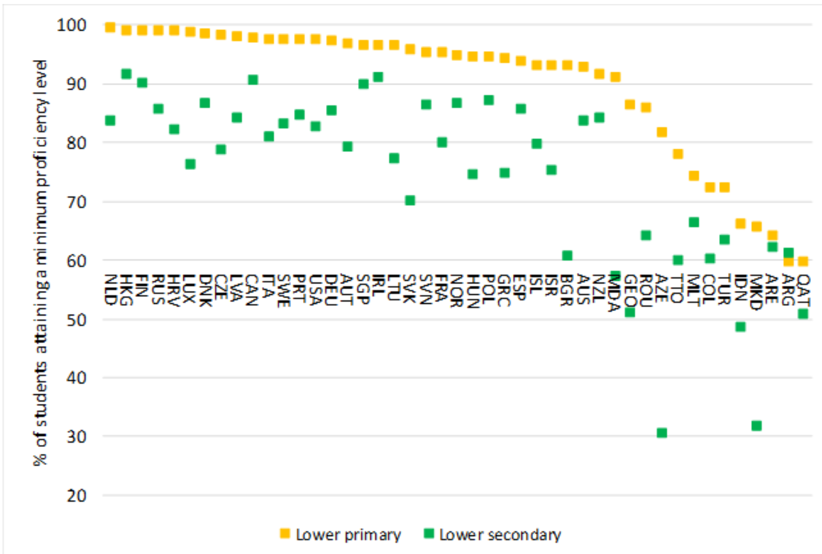
In constructing **Table 5** below, the most recent percentage proficient statistic per level and country found in the Altinok dataset was used. Only a few statistics from South Africa, Botswana and India were adjusted, in line with the discussion in section 7.2, before Table 5 was compiled. Only the minimum proficiency level (MPL) was considered. As will be discussed below, Altinok also calculated statistics for a basic minimum proficiency (BPL) level, but these are not used here. What stands out in Table 5 is that the higher the level, the *lower* the statistic of interest in general. For example, among the 65 countries in Altinok's data with mathematics statistics for both lower primary and lower secondary, the trend is for the statistic to be lower at the lower secondary level. At the lower primary level, the mean percentage proficient statistic across the 65 countries was 78% (no weighting by population used), while the figure was 73% for the lower secondary level. This would be in line with the notion that students fall increasingly behind. Some students do not cope at the primary level and fall behind stipulated standards. It is extremely difficult for these students to remedy their situation, and in each grade more students fall behind. One might expect the difference between the levels to be even greater when one takes into account that students drop out of schooling completely. This is explored in a subsequent section. All the six rows of Table 5 display the pattern of the statistic worsening the higher the level. Hence all values in the last column are negative.

Table 5: Differences across level in minimum proficiency level (MPL)

Countries	Lower prim.	End prim.	Lower sec.	Difference
READING				
5	52	38		-14
12		68	56	-12
45	90		74	-16
MATHEMATICS				
6	47	45		-2
16		45	43	-2
65	78		73	-5

Figure 7 illustrates country-level differences among the 45 countries with both lower primary and lower secondary reading statistics. The difference between the two levels varies from an average of 4 percentage points among the decile of countries with the smallest differences, to 39 percentage points among the top decile. Only one of the 45 countries, Argentina, displays the reverse: better proficiency levels at the lower secondary level than the lower primary level, though in Argentina the difference is a tiny one percentage point.

Figure 7: Reading MPL differences across level by country



The question is why the differences seen in Figure 7 emerge when Altinok's method is designed to produce similar proficiency levels at the lower primary and lower secondary levels (see section 6). These differences do not always emerge. If one compares mathematics, using just TIMSS results, across the two levels, one finds that among the 60 countries with this data, the average proficiency statistic for the lower secondary level is just 3.6 percentage points below that for lower primary. Just below half of the 60 countries have a lower secondary level proficiency statistic which is *higher* than the lower primary one. The patterns seen in Figure 7 all reflect situations where different assessment programmes apply to the two school levels. The explanation appears to be that the differences of Figure 7 are due to unintended effects of Altinok's linking process. These differences are what one might expect in schooling systems across the world, but Altinok's methodology was in fact aimed at removing such differences.

The obvious implication of the differences seen in Figure 7 for targets is that it is far more reasonable to set a 2030 target of 100% proficiency for the lower primary level than for the lower secondary level. To observe that the statistics tend to be worse at the secondary level, and then to conclude that it therefore is at *this* level where the greatest investments are needed, would almost certainly be a grave mistake. It would violate the notion that competencies at one grade are highly dependent on learning occurring at lower grades. Yet policymakers have been known to make such conclusions.

Table 6 indicates that where the two subjects are tested within the same programme, percentage proficient statistics (using MPL) are small. For instance, at the end of primary across 49 countries the average difference is 5 percentage points, with reading being higher. This is not surprising given that programmes tend to standardise the scores across the two subjects, and then apply similar or identical minimum thresholds, which were carried through into Altinok's calculations. However, even where different programmes apply, the difference is relatively small: 4 percentage points in favour of reading across 50 countries at the lower primary level. These 50 countries are all cases of TIMSS mathematics being compared to PIRLS reading. Having results from two different programmes does not appear to produce problematically large differences across subjects. Of course, this is largely by design. Though TIMSS and PIRLS are different programmes, they are both run by the IEA and follow similar methodologies. There are no instances in Altinok's data where the same school level has results across the two domains drawing from two completely unrelated testing programmes.

Table 6: Differences across subjects in minimum proficiency level (MPL)

	Lower prim.	End prim.	Lower sec.
READING MINUS MATH. – SAME PROGRAMME			
Difference		5	-4
No. countries		49	78
READING MINUS MATH. – DIFFERENT PROGRAMME			
Difference	3		
No. countries	50		

Note: Only the most recent statistic per country, level and subject was used.

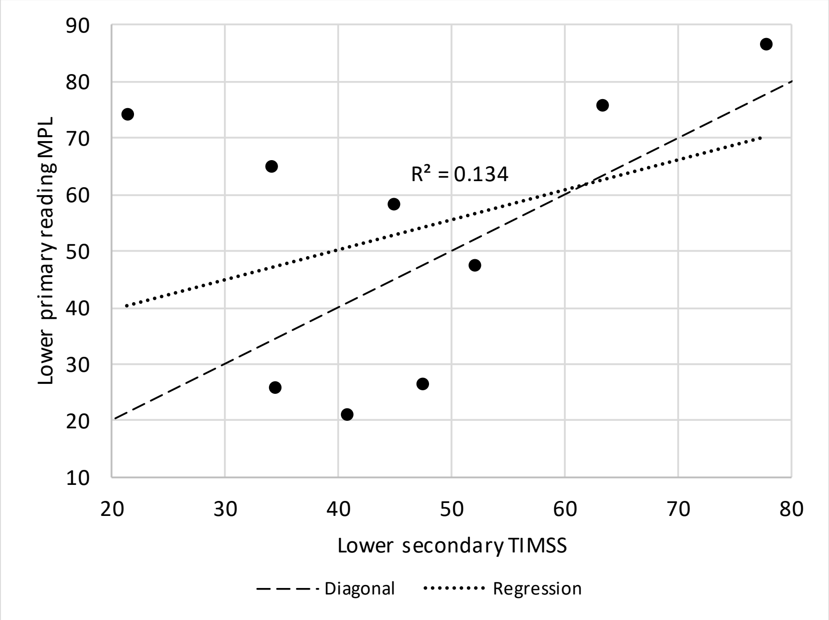
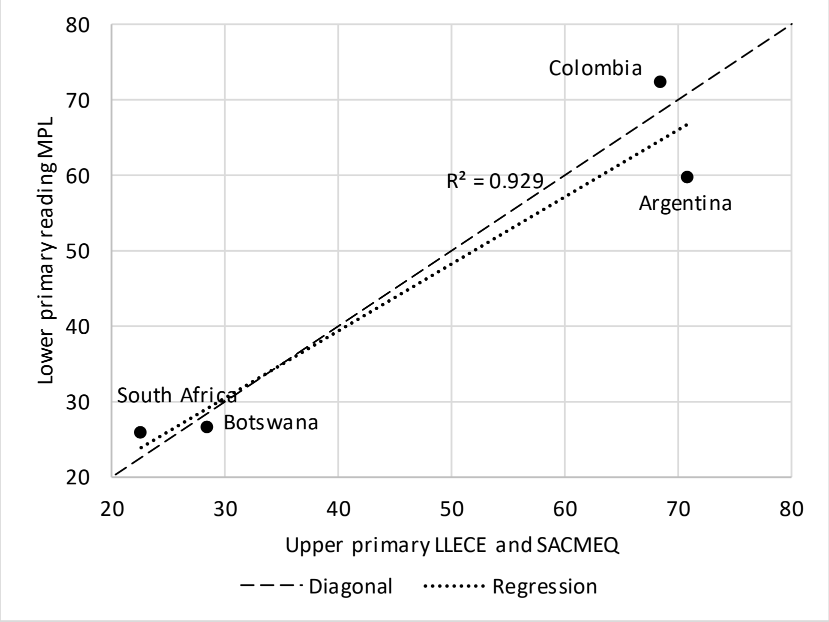
7.4 Imputation of missing percentage proficient statistics

This section describes how percentage proficient statistics not available in the 'original' Altinok dataset were imputed, either by using values for other subjects or levels found within the Altinok dataset, or through mostly World Bank GDP per capita statistics. 'Original' is in inverted commas because changes were made to the values of three countries, as explained in section 7.2. The process followed for lower primary reading, male and female combined, is explained in some detail. Thereafter, the key aspects of the imputation process for the remaining statistics are presented.

Whether the imputations could be implemented through a more automated approach than that described below was investigated. Specifically, standard methods existing for multiple imputations were tested. In the end, such methods were not workable, because the data being processed had too many missing values, and too few observations (or countries).

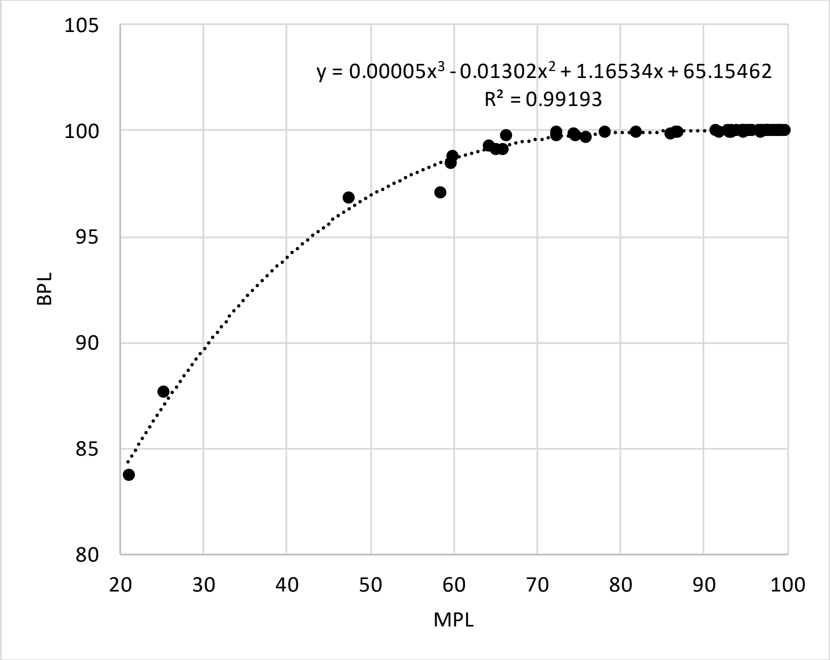
As mentioned in section 7.1, Altinok made use of only PIRLS as a source for lower primary reading. Of the 227 'countries' considered the universe, 54 had PIRLS-based lower primary reading proficiency statistics in Altinok's data. Thus values had to be imputed for the remaining 173 countries. Of these 173, 78 could receive imputations based on additional statistics in the Altinok dataset. The following four steps were followed in this regard, for the calculation of *minimum* proficiency level statistics:

Table 7: MPL imputations for lower primary reading

Steps	Imputed
1. A linear regression of PIRLS lower primary reading on TIMSS lower primary mathematics (50 observations, adjusted R^2 of .68) used for 16 countries with TIMSS but no PIRLS at lower primary.	16
2. A linear regression of PIRLS lower primary reading on PISA lower secondary both reading and mathematics (45 observations, R^2 of .62) used for 24 countries with PISA at lower secondary but no PIRLS at lower primary.	24
3. A linear regression of PIRLS lower primary reading on TIMSS lower secondary mathematics (9 observations) used for 5 countries with TIMSS at lower secondary but no PIRLS at lower primary. This was a problematic step, with R^2 (unadjusted) being just .13 (see below). However, the regression line seems intuitively right: a positive correlation, and lower primary exceeding lower secondary for much of the range. The five imputations were for: Ghana, Syria, Egypt, Palestine and Bosnia Herzegovina. 	5
4. A linear regression of PIRLS lower primary reading on LLECE and SACMEQ upper primary averages across reading and mathematics (4 observations after Honduras excluded) used for 33 countries with LLECE, SACMEQ or PASEC at upper primary but no value yet at lower primary. Few observations, but intuitively right, though lower primary will in many instances be lower than upper primary. 	33
Total	78

To impute statistics that use a *basic* proficiency level (BPL), the close relationship between the MPL and BPL statistics displayed in **Figure 8** was used. The cubic equation seen here was applied for all imputations of BPL, with respect to lower primary reading, even imputations occurring after MPL statistics were imputed using GDP per capita. (Points in Figure 8 are countries with both values present before any imputation.)

Figure 8: BPL and MPL



Throughout, any imputed MPL or BPL values exceeding 100 were reduced to 100.

Imputations drawing only from the Altinok dataset resulted in 132 countries with lower primary reading values, meaning a further 94 countries needed imputed values drawing from external economic data. These 94 represent 21% of the world's child population. World Bank values for 'GDP per capita, PPP (current international $)' were used[42]. For 59 of the 94 countries, World Bank figures for 2017 were available. For the remaining 35, virtually all very small 'countries', World Bank values from before 2017 or values from some alternative source were used. Lower primary reading MPL statistics were regressed on the natural log of GDP per capita and the UNESCO world region dummies, resulting in an adjusted R^2 of .74. This regression was used to predict MPL values for the remaining 94 countries, and thereafter BPL values were calculated as explained above.

Table 8 provides a crucial reality check of the statistics obtained. This table first reproduces values found in Altinok (2017: 49). These values refer to primary schooling as a whole, not lower primary. It was possible to roughly reproduce these values using the original Altinok dataset. One reason it would not be exact is that Altinok does not appear to explain from which years he drew the data, how he weighted countries, and how he combined lower and upper primary. However, the rough equivalence is reassuring. Differences would in

[42] World Bank Open Data at https://data.worldbank.org, accessed June 2019.

part be due to the difference between the world regions used by Altinok, and the UNESCO regions used for the current report.

The differences between the MPL ('Min.') values before and after imputations warrant discussion. Imputations make the Central and Southern Asia value drop from 76% to 54%. This is mainly due to the inclusion of India. Likewise, the increase for Eastern and South-eastern Asia, from 67% to 86%, is mainly due to the inclusion of China.

The final bottom-right figure in Table 8 of 59% means this that worldwide 41% percentage of students at the lower primary level in the world do not reach the level of reading proficiency they should.

Table 8: Comparing Altinok's statistics to new lower primary reading statistics

All primary						*Lower* primary only					
	Altinok (2017) Figure 2			Altinok replicated		From just 'original' Altinok		New after Altinok-only imputations		New after GDP per capita imputations	
	Min	Basic		Min	Basic	Min	Basic	Min	Basic	Min	Basic
SSAfrica	15	63	SSAfrica	15	67	26	87	20	83	20	83
Arab States	50	88	N. Africa & W. Asia	60	96	61	96	54	96	54	96
S. & W. Asia	76	100	Central & Southern Asia	64	99	76	100	54	98	56	98
Central Asia	84	100									
East Asia & the Pac.	91	100	Eastern & S-Eastern Asia	87	100	67	100	86	100	85	100
			Oceania	92	100	93	100	93	100	83	100
Latin Am. &Car.	73	99	Latin Am. & Car	66	97	67	99	69	99	68	99
North Am. & W. Europe	96	100	Europe and North Am.	94	100	97	100	96	100	96	100
			World	62	90	79	99	64	96	59	95

7.5 Taking into account out-of-school children and adolescents

This section describes how national statistics on participation in schooling were used to derive the percentage of the *population* becoming proficient. Though there is some ambiguity in the documentation around whether indicator 4.1.1 is concerned with just students, or the whole population of school-age children, there appears to be general agreement that out-of-school children do need to be taken into account.

For lower primary, one can think of the out-of-school as the percentage of children who will reach the required proficiency level associated with the lower primary level by the time they have reached the age at which children typically complete this level. It is not possible to precisely calculate this based on the national statistics that are easily available. For instance, no tables exist with, per country, the percentage of children aged nine at the start of the calendar year who are not attending school.

Household-derived data on the percentage of children not in school are clearly more reliable than calculations involving the comparison of Ministry of Education enrolment data to official population estimates. There is clear evidence that this is the case – see for instance Stukel and Feroz-Zada (2010). Two sets of household-derived statistics available in UIS.Stat were used[43]: 'Rate of out-of-school children of primary school age, both sexes (household survey data)' and 'Completion rate, primary education, both sexes (%)'. Both statistics can be used to approximate the ideal, which would be the percentage of children of an age cohort, around age nine, who are not in school. The first would be the same as the ideal if the population was of a constant size across ages and children either participated fully in primary schooling, or not at all. One hundred per cent minus the second statistic would also be the same as the ideal if these two conditions held. The two statistics were present for 89 and 92 countries respectively – 89 had both statistics, and these 89 countries represented 67% of the world's population of children. The great majority of the 92 countries were developing countries. Only two countries from the Europe and Northern America region, namely the United States and Montenegro, had statistics.

The population-weighted mean for the out-of-school rate, for 89 countries, was 13%, while that for the completion rate was 81%, meaning 19% of children are not completing primary education. In 70 of the 89 countries, the out-of-school rate produced the lowest *non*-participation value, as one might expect. Among the 92 countries, whichever of the two statistics – the out-of-school rate or the completion rate – produced the lower out-of-school value, was assumed to offer the best basis for children aged around nine not in school. This new best statistic produced an overall mean of 12% for the 92 countries. India's final statistic was 5%. Statistics for China were missing from the UIS.Stat tables.

An OLS regression represented by the following was run to impute out-of-school values for countries with no statistics. The natural log of the out-of-school statistic described above was regressed on the log of GDP per capita (see discussion in section 7.4) and 0-1 dummies for sub-Saharan Africa as well as Europe and Northern America. These two regions clearly differed from the global norm, with the out-of-school figures being higher in sub-Saharan Africa and lower in Europe and Northern America. R^2 was .66.

$$\ln(OOS) = \beta_0 + \beta_1 \ln(GDPCap) + \beta_2 SSA + \beta_3 ENA \qquad (1)$$

The imputed out-of-school value for China was 1.5%, which seemed plausible. This is slightly better than the non-imputed 2.0% for Viet Nam or 1.9% for Thailand. With all 227 countries carrying out-of-school statistics, a global population-weighted mean of 9% emerged. This 9% would be within the range of estimates reflected in UIS (2017a: 5, 11): 3% of primary-level children never enter school, while 18% of children do not complete primary schooling.

In line with UIS (2010b: 9), it was assumed that children not in school would not have reached either a minimum or a basic level of proficiency. The following equation was applied to arrive at a percentage of the lower primary-aged *population* not being proficient:

$$P = S \times (1 - N) \qquad (2)$$

[43] Accessed June 2019.

The percentage proficient among enrolled students (*S*) is multiplied by one minus the percentage of the population not in school (*N*). The equation was applied for both the minimum and basic proficiency statistics.

The first two columns of statistics in **Table 9** are the last two columns of Table 8. The last two columns reflect values after the out-of-school phenomenon has been taken into account. The global percentage reaching a minimum level of reading drops from 59% to 56%, largely due to changes in the first three world regions listed in the table. This 56% is higher than the 44% figure reflected in UIS (2017a: 5) for primary schooling *as a whole*[44]. As discussed previously, one might expect a higher value at the lower primary level compared to primary schooling in general. The picture seen in Table 9 is roughly compatible with the narrative already created that 'More than one-half of children and adolescents are not learning worldwide' – this is from the title of UIS (2017a). While 56% for lower primary is more than half, taking into account older children (as explained in section 7.6), lowers the global figure to 50%. However, that is if one considers the minimum proficiency level, or MPL. If one instead looks at 'Basic', implying 12% of the world's lower primary-aged children are not becoming proficient, a completely different understanding emerges. The MPL baseline scenario appears to be the more useful one. Not only is it in line with understandings in the UIS, even national governments seem to accept this as being a reality on which to base policies and targets[45].

Table 9: Lower primary reading proficiency in population

	% of students proficient		% of a cohort of primary-aged population that is proficient	
	Min.	Basic	Min.	Basic
Sub-Saharan Africa	20	83	16	64
Northern Africa and Western Asia	54	96	50	88
Central and Southern Asia	56	98	50	88
Eastern and South-eastern Asia	85	100	83	97
Oceania	83	100	82	97
Latin America and the Caribbean	68	99	67	97
Europe and Northern America	96	100	95	100
World	59	95	56	87

7.6 Adjustments and imputations for remaining dimensions

The first step taken in dealing with dimensions other than lower primary reading, both sexes, was to obtain best possible values for all countries on the percentage of adolescents of the age usually associated with the end of lower secondary schooling, who were not in school. Procedures (and challenges) similar to those relating to the out-of-school at the primary level (see section 7.5) applied. Two UIS.Stata datasets were used: 'Rate of out-of-school adolescents of lower secondary school age, both sexes (household survey data) (%)'

[44] This source states that 55% of females and 57% of males at the primary level do not reach a minimum level of proficiency, giving an average of 56% of children *not* achieving proficiency.
[45] To illustrate, low proficiency baseline statistics in Botswana's (2015) education sector plan would be in line with the 400 TIMSS minimum level.

and 'Completion rate for lower secondary education (household survey data)'. Combined, they provided values for 119 countries. Income per capita values were used to impute values for the remaining countries. The end result was a 14% population-weighted out-of-school percentage for the world. India's figure was 9%, and China's 17%. Both of these values are non-imputed, although India's draws from the out-of-school variable, while China's draws from the completion rate variable. China's out-of-school value was missing. In line with what was done at the primary level, whichever of the two variables yielded the lowest out-of-school figure was used. Had India's out-of-school value been missing, and had India also relied on the completion variable, India's final value would have been close to China's, at 19%.

The next step was to adjust all proficiency statistics in the Altinok dataset, other than the already completed lower primary reading statistics, using out-of-school figures, and equation (2). For lower primary mathematics and end of primary, the primary level out-of-school values were used, and for end of lower secondary, the out-of-school values derived for that level were used. **Table 10** provides aggregate statistics emerging from a total of 133 countries with values in the Altinok dataset, after the corrections described in section 7.2 and after the out-of-school adjustment, but before the imputation of values for countries with missing data in the Altinok data. Though lower primary reading values are not the focus here, the corresponding values for that category are provided for the sake of completeness.

Table 10: Attainment of MPL in population before imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	25	38	11	18	54	27
Northern Africa and Western Asia	60	50			47	46
Central and Southern Asia	75	71			18	22
Eastern and South-eastern Asia	66	55			64	66
Oceania	92	89			83	87
Latin America and the Caribbean	65	42	76	54	55	44
Europe and Northern America	97	93			81	87
World	78	69	36	32	50	50

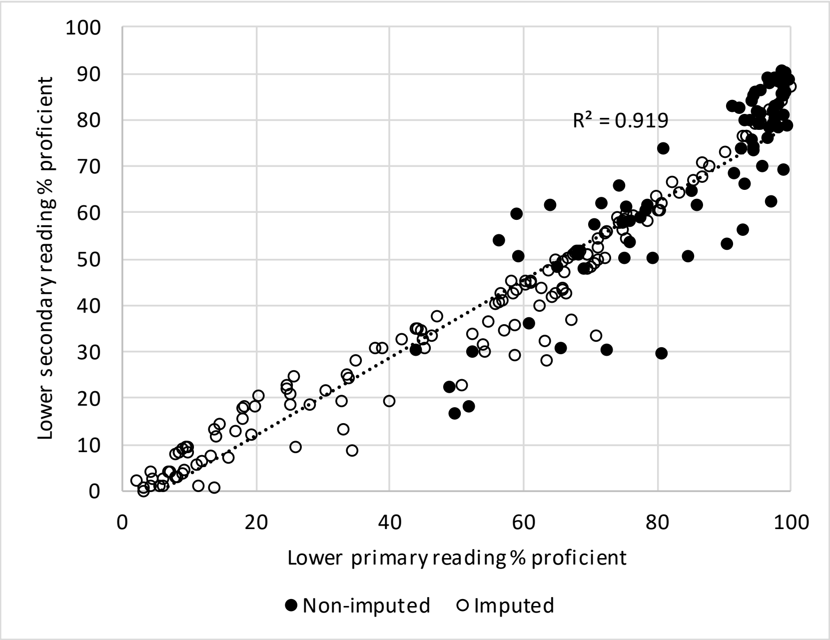
A general approach for the imputation process was then considered. The conclusion was that it would be optimal to pay special attention to maintaining as far as possible credible relationships across the three levels, and across the two domains. These relationships are important to understand if planning and targeting of SDG 4.1.1 is to proceed in an informed manner, both globally and at a national level. The relationships were introduced in section 7.3. In the steps explained below, the lower primary reading statistics already derived for all 227 countries are an important anchor for other statistics.

The following emerged as an efficient predictor of lower secondary reading proficiency percentages, considering an age cohort, not just enrolled students. In particular, equation (3) ensured that no developing countries ended up with a left-hand value below zero. The three coefficients 19.11, 0.015 and -0.022 were obtained through a nonlinear regression involving 45 countries which had values in the original Altinok dataset at both the lower primary and lower secondary levels, for reading. P_{LP} is the percentage of the population proficient in reading at the lower primary level, and OOS_{LS} is the out-of-school percentage for the lower secondary level discussed above. *e* is the base of the natural logarithm.

$$P_{LS} = 19.11 \times e^{(0.015 \times P_{LP} - 0.022 \times OOS_{LS})} \quad (3)$$

Equation (3) was used to obtain P_{LS} for the 227 countries. In the case of 32 countries where imputed P_{LS} exceeded the corresponding value at the lower primary level, P_{LS} was made to equal the lower primary value. Similarly, in the case of 26 countries where imputed P_{LS} exceeded the upper primary reading value, P_{LS} was made to equal the upper primary value. It was thus assumed that in any education system, attainment of the minimum proficiency would decline as one moved up the system, an assumption supported strongly by the earlier Figure 7. **Figure 9** below illustrates the percentage of the population, at the two levels, who were found to be proficient. 'Non-imputed' here means that the lower secondary value existed in the Altinok data. Clearly, there was a lack of data at the lower secondary level for the least developed countries.

Figure 9: Lower primary and lower secondary proficiency



For upper primary reading statistics, the point of departure was weak, given that only two of seven world regions had any data (see Table 2). Where the data existed for both primary levels, proficiency levels were on average slightly higher, but by just one percentage point on average, at the upper primary level compared to the lower primary level, and considering reading. In the absence of further information, the most defensible seemed to be to repeat the lower primary value at the upper primary level, where the latter was missing.

Turning to mathematics, one would not expect very large differences between reading and mathematics statistics, given that the testing systems that test both subjects aim to produce the same mean across the two. However, there are differences, driven to a large extent by situations where different testing systems test the two subjects in the same country. The most defensible seemed to respect reading-mathematics differences seen in the original Altinok data, and to replicate these where imputations occurred. For the lower primary and lower secondary levels, country mathematics scores were regressed on country reading scores, where both sets of scores existed in the original Altinok data. The coefficients for the regression were then used to impute values. All imputed values fell within the acceptable range of 0 to 100. For the upper primary level, once again the approach was taken of using the lower primary value, where no original value existed.

The outcome of the imputations is shown in **Table 11**. If one compares Table 11 to Table 10, global values drop for the lower primary and lower secondary levels. This is because more developing countries are brought into the calculation. In contrast, the upper primary values rise because in the original dataset all countries with data are developing countries. The mean across the three reading columns, at the global level, is 50%. Thus, the overall picture is that around half of children and adolescents are not reaching a minimum proficiency level.

Table 11: Attainment of MPL in population after imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	16	25	14	21	10	13
Northern Africa and Western Asia	50	51	50	51	34	40
Central and Southern Asia	50	53	50	53	21	24
Eastern and South-eastern Asia	83	77	83	77	60	64
Oceania	82	80	82	80	67	71
Latin America and the Caribbean	67	63	77	54	51	43
Europe and Northern America	95	92	95	92	80	86
World	56	57	56	55	37	40

The next step involved imputing gender-specific values where these were missing. **Table 12** below displays the breakdown of the 115 countries in the Altinok data with gender-specific proficiency statistics. Crucially, both India and China lacked such statistics, even at the lower secondary level. This clearly creates problems of reliability in relation to gender, beyond the problems already outlined above relating to the fact that these two countries, accounting for one-third of the world's children, have no test data at all at the primary level (at least not data which were easily available for the current study).

Table 12: Breakdown of 115 countries with MPL by gender values

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	2	2	29	30	1	3
Northern Africa and Western Asia	10	16			9	15
Central and Southern Asia	1	2			1	2
Eastern and South-eastern Asia	3	8			6	8
Oceania	2	2			2	2
Latin America and the Caribbean	5	4	19	19	12	14
Europe and Northern America	32	32			34	36
World	55	66	48	49	65	80

The 115 countries produce the gender parity index (GPI) values presented in **Table 13** below. For each country, school level and subject, the proficiency statistic for females produced by Altinok was divided by that for males. Some of the statistics are extraordinary, indicating for instance that in the Northern Africa and Western Asia region, 44% more females than males enrolled at the lower secondary level are proficient in reading. Differences by gender with respect to percentage proficient statistics are often not published. Notably, such differences are not included in the main TIMSS and PIRLS reports. The 2015 TIMSS Grade 4 mathematics microdata, downloadable through the TIMSS-PIRLS website, were analysed to verify the patterns. Countries with the highest GPI values for their percentage proficient statistics, using a threshold of 400 TIMSS points, are shown in **Table 14**. Clearly, already at the lower primary level, several countries in the Northern Africa and Western Asia region display remarkably high values for females relative to males. (Unusually, Altinok [2017: 50] calculates the gender parity index by dividing male values by female values. This explains why Altinok's values are roughly the reverse of the values seen in Table 13.)

Table 13: Gender parity in the test data before imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	1.20	1.16	1.00	1.10	1.33	1.09
Northern Africa and Western Asia	1.26	1.10			1.44	1.07
Central and Southern Asia	1.11	1.08			1.15	1.04
Eastern and South-eastern Asia	1.15	1.11			1.26	1.09
Oceania	1.04	1.00			1.12	1.01
Latin America and the Caribbean	1.07	0.86	1.04	0.94	1.17	0.88
Europe and Northern America	1.01	1.00			1.10	1.00
World	1.09	1.05	1.02	1.04	1.19	1.01

Table 14: Highest GPI values in TIMSS Grade 4 2015

	Females	Males	GPI
Saudi Arabia	53	33	1.60
Kuwait	35	30	1.17
Oman	65	56	1.17
Indonesia	49	43	1.13
Bahrain	77	69	1.12
Iran	64	58	1.11

What is also striking about the Table 13 statistics is that in the Latin America and the Caribbean region, female performance in mathematics emerges as relatively weak. This pattern is unique to this region, and is seen at all three levels of the schooling system.

Table 15 displays GPI values in relation to in-school statistics, which are the complements of the out-of-school statistics discussed above. For the primary level, one statistic per country applies, and for the secondary level another statistic. Differences within rows are thus in part due to which countries, of the 115, participated in which tests. Though all the 115 countries have original proficiency statistics in the Altinok dataset, some out-of-school statistics are imputed. Specifically, of the 115 countries, 86 countries had non-imputed primary out-of-school statistics, while all 115 had non-imputed statistics for the lower secondary level. Deviations from 1.00 in Table 15 are relatively small, pointing to relatively high levels of gender parity when it comes to school participation (at least in the 115 countries).

Table 15: Gender parity in the participation data before imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	1.00	1.00	0.99	0.99	1.01	1.00
Northern Africa and Western Asia	1.00	0.98			0.99	1.00
Central and Southern Asia	1.00	1.00			1.01	1.01
Eastern and South-eastern Asia	1.00	1.00			1.00	1.01
Oceania	1.00	1.00			1.01	1.01
Latin America and the Caribbean	1.01	1.01	1.00	1.00	1.00	1.01
Europe and Northern America	1.00	1.00			1.00	1.00
World	1.00	1.00	0.99	1.00	1.00	1.00

Table 16 shows the net GPI after taking into account performance in assessments and school participation. Even after small disadvantages for females with respect to participation are taken into account (for instance 0.99 for the lower secondary level in reading in Northern Africa and Western Asia), the overall picture of a clear educational advantage for females remains.

Table 16: Net gender parity before imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	1.21	1.16	1.03	1.10	1.34	1.10
Northern Africa and Western Asia	1.25	1.08			1.42	1.07
Central and Southern Asia	1.11	1.08			1.16	1.05
Eastern and South-eastern Asia	1.15	1.11			1.26	1.10
Oceania	1.04	1.01			1.13	1.01
Latin America and the Caribbean	1.08	0.87	1.04	0.94	1.18	0.88
Europe and Northern America	1.01	1.00			1.10	1.00
World	1.10	1.05	1.04	1.04	1.19	1.01

In attempting to impute gender-specific values where these were missing, conditional correlations between income per capita, the level of proficiency, region and female advantage or disadvantage were explored. This did not reveal clear and consistent patterns. The decision was thus taken to apply the GPIs displayed in Table 16 to the already calculated two-gender statistics, where imputations were needed. The aggregate statistics, with all 227 countries having gender values, are shown in the next three tables. The means of the proficiency statistics, for females and males, are shown in the last two tables.

Table 17: Net gender parity after imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	1.25	1.16	1.13	1.13	1.34	1.11
Northern Africa and Western Asia	1.25	1.08	1.25	1.08	1.46	1.08
Central and Southern Asia	1.11	1.05	1.11	1.05	1.16	1.04
Eastern and South-eastern Asia	1.08	1.07	1.08	1.07	1.19	1.07
Oceania	1.04	1.01	1.04	1.01	1.13	1.02
Latin America and the Caribbean	1.14	0.89	1.05	0.94	1.18	0.88
Europe and Northern America	1.02	1.00	1.02	1.00	1.11	1.00
World	1.14	1.07	1.11	1.06	1.23	1.05

Table 18: Attainment of MPL in female population after imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	18	27	15	23	11	14
Northern Africa and Western Asia	55	53	55	53	40	41
Central and Southern Asia	53	54	53	54	23	24
Eastern and South-eastern Asia	86	79	86	79	65	66
Oceania	83	80	83	80	71	71
Latin America and the Caribbean	71	60	77	52	55	40
Europe and Northern America	96	92	96	92	84	86
World	59	58	59	57	40	41

Table 19: Attainment of MPL in male population after imputations

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	14	23	15	20	8	12
Northern Africa and Western Asia	45	49	55	49	28	39
Central and Southern Asia	47	52	53	52	20	23
Eastern and South-eastern Asia	80	74	86	74	55	62
Oceania	80	79	83	79	63	70
Latin America and the Caribbean	62	67	77	56	47	45
Europe and Northern America	94	92	96	92	76	86
World	53	56	59	54	34	39

8 Population projections

Where population figures were used in the analysis above, these were largely historical figures obtained from UIS.Stat (see the discussion preceding Table 1). For the projections of the global reach of learning proficiency, however, population projections would be necessary. In particular, a concern one may have is that globally, progress in learning proficiency could be slowed down by the fact that countries with a lower point of departure display higher population growth.

Clearly, the best available source for population projections is the 2019 World Population Prospects of the United Nations Population Division (UNPD). The 'medium variant' projections were selected – there is also a 'high variant' and a 'low variant'[46]. The medium variant is often the preferred variant used in United Nations reports[47].

The UNPD projections provide a value for age categories consisting of five years, and for every five years in time, for 2020 to 2100. The age category 5 to 9 was selected to represent one age cohort for all dimensions of indicator 4.1.1. The complexity of considering the age 10 to 14, or even 15 to 19, category for lower secondary was avoided, largely because this would not make a significant different to the *trends.* Moreover, separate statistics for male and female were used as this seemed important. A lower number of females than males, aged 5 to 9, is common: in 2020, in around one-quarter of countries, females come to less than 95% of males. Moreover, this anomaly is expected to decline slightly over time, with the across-country average female-to-male ratio rising from 93% to 95% between 2020 and 2100.

For 58 countries, all countries with very small populations, population figures did not exist in the UNPD source and were therefore sought elsewhere, for instance in UIS.Stat. Thereafter, baseline values for 2015 were imputed based on a quadratic regression of years on population for the years 2020 to 2035.

9 The structure of the Excel tool

Projections of distributions of learning outcomes, and of the attainment of specific proficiency levels, are implemented in the Excel tool.The structure of the tool is explained, in broad terms, below, following the order of the various sections in the tool. Inspection of the actual tool, which has no hidden or locked elements, would reveal further details.

Inputs on rate of progress are about the assumed standard deviations a year by which the mean scores of countries progress. First a 'speed limit' is specified, based on research that accompanies the current report. This speed limit differs by the score at the point of departure, with larger improvements being more possible the lower the starting point. Secondly, an assumption is made, per world region, of the percentage of the speed limit that will be attained. The default is a speed limit ranging linearly from 0.075 standard deviations a year for a country with a starting score of 250, to 0.041 standard deviations a year for a country with a starting score of 550. This realises a speed limit derived from a separate analysis of PIRLS trends, specifically an annual improvement of around 0.07 for a country such as Morocco or South Africa, and around 0.04 for a country like Russia. If one assumed that, say, sub-Saharan Africa was to achieve 50% of the speed limit, each country in this region would advance half as much between a base year and the target year (assumed to be 2030, but this can change) as it would have advanced if it had reached the speed limit. If the intention is to produce a business-as-usual (BAU) set of projections, then BAU parameters would be inserted as the 'speed limit' and all world regions would be set to achieve 100% of this.

Inputs on standard deviation. It is common to adopt the rough assumption that the standard deviation in each and every country is 100 in testing programmes using the '500/100' metric, meaning a mean across all students across all countries (in some base year) of 500, and a standard deviation across all these students of 100. Altinok's (2017: 18, 38) scale is this '500/100' scale, specifically that of the TIMSS and PIRLS scales. The tool permits the use of the assumption that all countries display the same standard deviation of 100. To do

[46] Excel files for 'Population by Age Groups – Female' and 'Population by Age Groups – Male' were downloaded off https://population.un.org/wpp/Download/Standard/Population (July 2019).
[47] See for instance United Nations: Department of Economic and Social Affairs, 2019.

this, the user would insert 100 into both standard deviation cells. These are the default values for the tool. However, it is also possible to change this and assume that the standard deviation, which reflects educational inequality, declines as countries progress. Such a trend is seen in the actual data insofar as countries with lower scores in PIRLS (and PISA) display a lower standard deviation. This is discussed in the accompanying report. Specifically, PIRLS suggests a differentiated standard deviation prevails where a score of 250 is associated with a standard deviation of 120, and a score of 550 is associated with a lower standard deviation of around 72. The effect on the projections of inserting these parameters is discussed below.

The reason for not inserting into the tool the actual and empirically derived standard deviation for every country is that this information is not available to the necessary degree. In particular, they are not included within the Altinok data discussed in section 7.1. While standard deviations are published in, for instance, the official TIMSS reports, not all countries have such published standard deviations. Imputing baseline standard deviations for countries without published figures, on top of the imputations of the percentage proficient statistics (see section 7.4), would have made the whole projection process far more complex, and the benefits of this would have been difficult to gauge.

Choosing the proficiency thresholds is largely about choosing whether to use a basic level of proficiency (BPL), or a minimum level of proficiency (MPL), for the projections. This differentiation follows Altinok (2017: 62). The defaults are a score of 400 for the MPL and 250 for the BPL. However, BPL as an alternative is only available for one indicator, namely the one for lower primary reading. Given the discussion preceding Table 9, it was decided it was not worth including any further BPL sets of values in the tool.

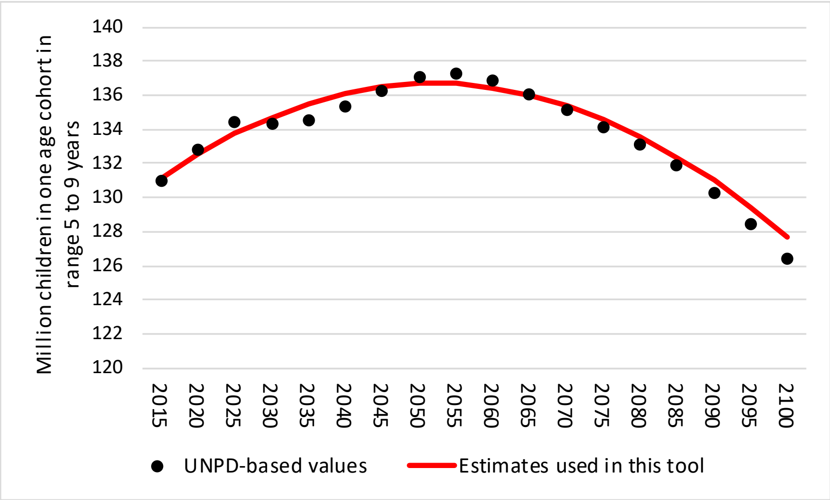
A **baseline year** is selected, for instance 2015. For the scenarios described below, 2015 was selected as this is the year in which the SDGs were adopted. As seen from Table 4, this implies using performance levels from on average around 2010 to represent the situation in 2015.

Choosing what to graph involves selecting which of the 19 dimensions of the SDG indicator to use.

Input on population projections involves specifying a future outer year defining the range for which country-specific population trendlines are determined. The UNPD-based population values have a value only for every fifth year. It seemed important to allow for any future year to be selected as a year for which a future scenario could be generated, not just years which are multiples of five. This meant values for years not in the UNPD series had to be estimated. One approach would have been to assume a linear trend for the intervening years. However, overall trends are unlikely to follow this pattern. Instead, a quadratic regression of UNPD-based population values on year and year squared seemed preferable, as this would produce a smoother and more plausible trend.

The input required for the tool is the outer year for the series of data used in the regression. For example, if the outer year is 2070, then the regression is run (within the tool, using Excel's matrix algebra array functions) using values for the period 2015 to 2070. The tool contains a graph which illustrates, at the global level, the deviation between predicted values, which are the values used for the projections, and the original UNDP-based values. Error! Reference source not found. illustrates the gap between the two, where the selected outer year is 2075. Note that predictions occur even beyond 2075, though the coefficients use only 2015 to 2075 data. Arguably, the resultant red curve presents a trend which is at least as plausible as the original UNPD-based one. Specifically, the latter displays an atypical decline between 2025 and 2030, which is not easily explained. Closer examination of the data indicates that this 'kink' in the trend is driven largely by the North Africa and Western Asia region, and in particular a 'kink' in Egypt's trend.

Figure 10: Size of one global age cohort 2015 to 2100



The tool generates three **graphs**. All three illustrate distributions of scores. The first shows the situation in the baseline year, using the assumption that the country-specific input data represent sufficiently well the picture in the baseline year. The second graph does the same for the target year, which should probably always be 2030, the horizon year for the SDGs. The first two graphs break down values by world region, with India and China represented separately. The third graph contains four global distributions, for the baseline and target years, and two additional user-selected years, which can be beyond the target year. The baseline, target and two additional years are each associated with a global percentage proficient statistic.

Non-changing country-level input data on proficiency appearing in the tool consists of the percentage of children who reach specific proficiency levels. There are 19 dimensions of indicator 4.1.1 in the input data, 18 for the MPL, and one (largely for demonstration purposes) for the BPL. The 18 MPL indicators are defined by the three school levels, two subjects and two sexes.

Non-changing country-level input data on population change would be UNPD-based projections, as described in section 8 above, with female and male totals presented separately. For each country, there are three regression coefficients for females, and the same for males, which are calculated within the tool (so they can be sensitive to the outer year parameter described above). These coefficients are used to calculate values per year used in the tool, including values for years which are not multiples of five. The following equation would be used, where P is the age 5 to 9 population total and Y is the year and a, b and c are the three coefficients.

$$P = a + bY + cY^2 \tag{4}$$

The **background calculations** of the tool are fairly self-explanatory.

The next two graphs (which use data from the tool, but are not found in the tool) help illustrate the approach. **Figure 11** displays normal curves per country, at least for countries with relatively large populations. The situation for lower primary reading, using MPL, and including both male and female, applies. The vertical axis must be explained. To illustrate, the value for India at score 375 is 2.4. This means there are 2.4 million children of an age cohort, of around age nine, who, when their score is rounded to the nearest multiple of 25, end up with 375. Points that produce the curves are set at each multiple of 25.

Figure 11: Baseline MPL-based patterns by country

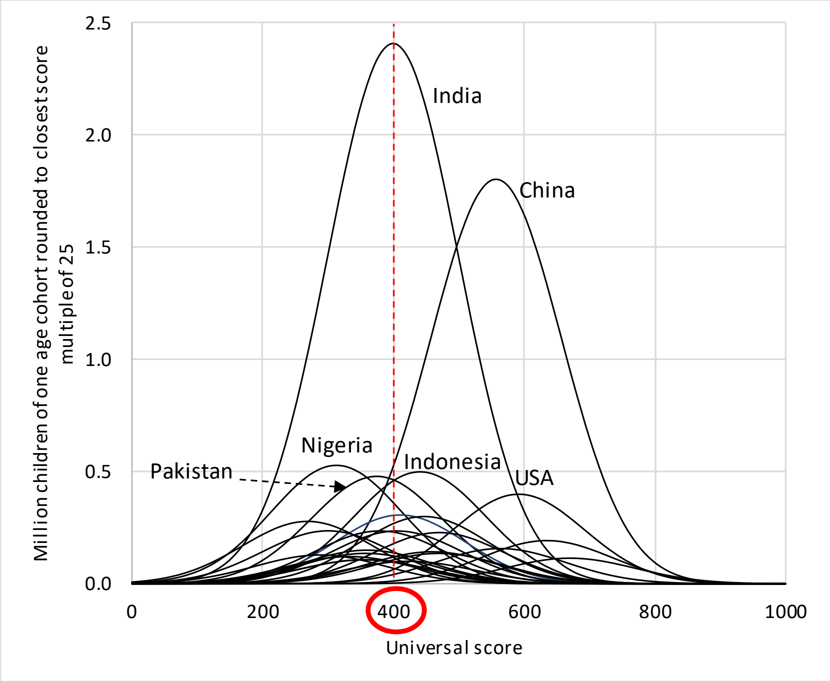


Figure 12 displays world regions, which are stacked national curves. What is remarkable is that each regional curve appears strongly normal. In fact, skewness and kurtosis statistics confirm that the distributions are about as normal as the TIMSS country distributions seen in earlier Figure 1. Regions are thus remarkably homogenous as far as learning outcomes are concerned. Figure 13 displays what one does *not* find. In **Figure 13**, the regions 'Europe and Northern America' and 'sub-Saharan Africa' have been combined. Had world regions been highly unequal, non-normal distributions of this type would have emerged.

Figure 12: Baseline MPL-based patterns by world region

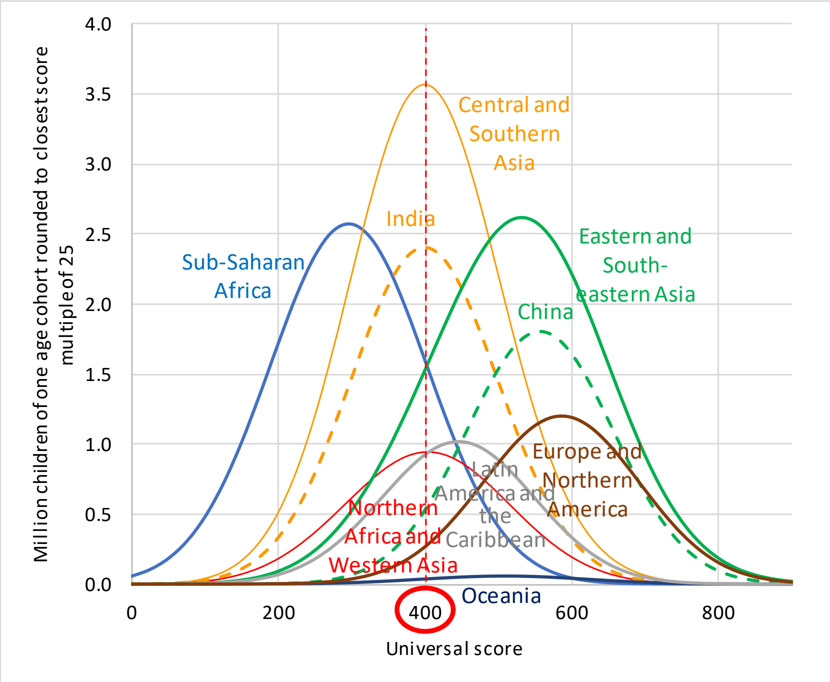
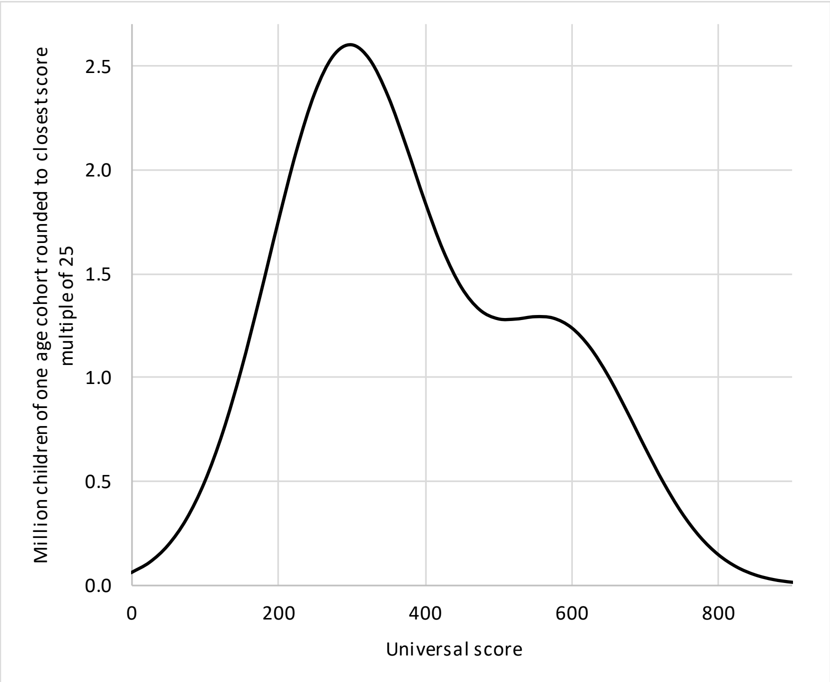


Figure 13: A hypothetical combination of two world regions



In the next two graphs, which are found in the tool, regions are stacked on top of each other, with India and China marked separately. Here again, lower primary reading is illustrated. Even globally, the distribution of learning outcomes, or scores, is normal, both in terms of its visual appearance and its normality statistics.

The baseline seen in **Figure 14** represents 57% of children reaching the MPL of 400 (the reason the corresponding figure is 56% in the previous Table 9 is that Table 9 uses a different source for population data).

Figure 15 represents a 2030 scenario, where 75% of children have reached the MPL. This scenario is optimistic in the extreme, as it assumes that the entire world would progress in line with the default 'speed limits' described above – an annual improvement of 0.075 standard deviations for countries such as Morocco and South Africa, and 0.041 for Russia.

Figure 14: Baseline MPL-based patterns by world region (from tool)

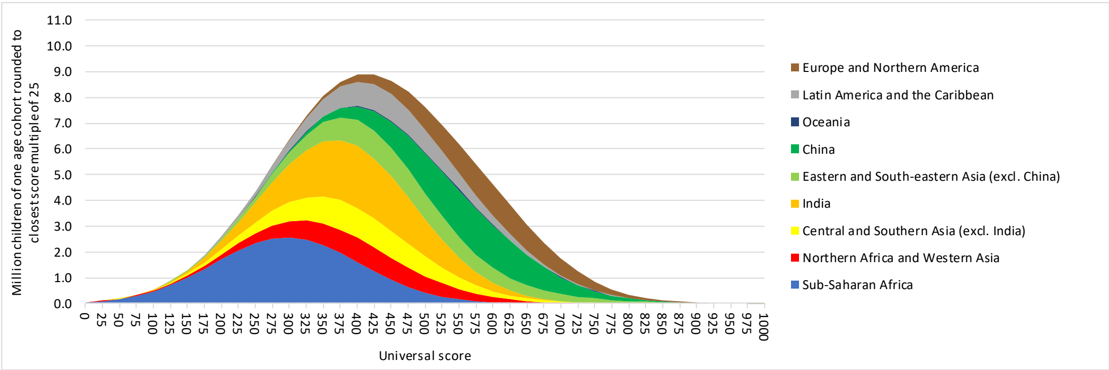


Figure 15: Target year MPL-based patterns by world region (fast progress)

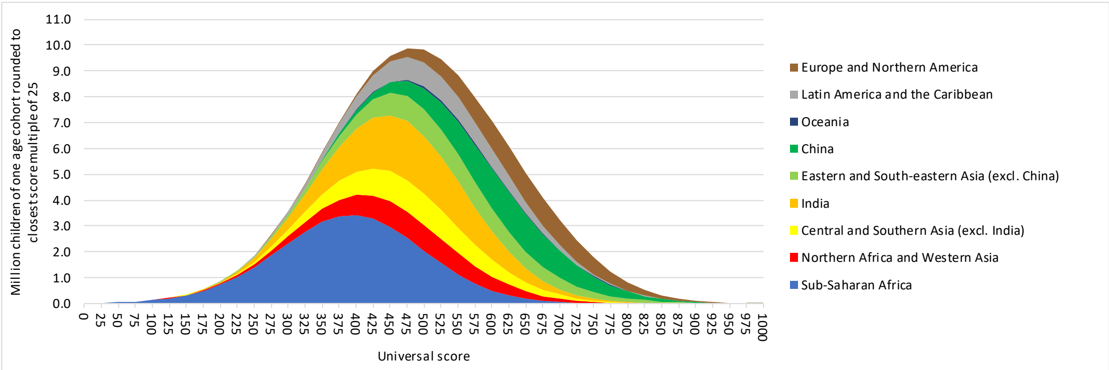
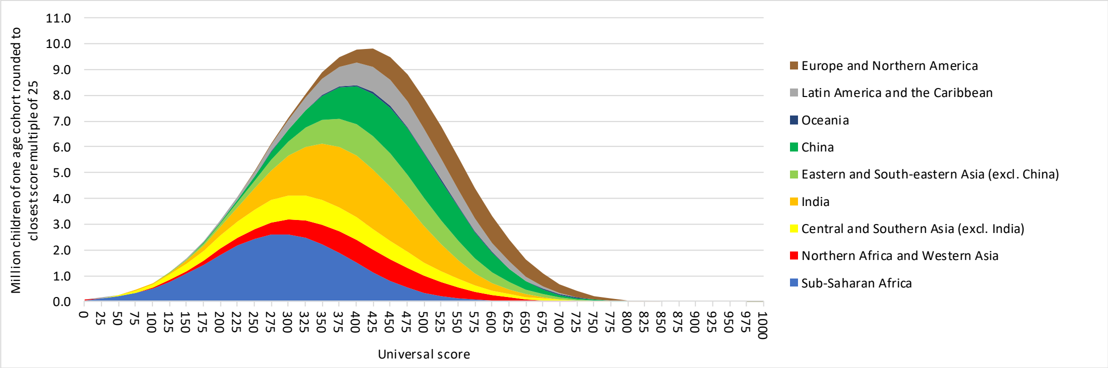


Figure 16 represents the baseline if the user instruction specifies that the basic proficiency level (BPL) should be used. This should produce a graph that is close to Figure 14, the MPL-based baseline, though it would not be exactly the same. This is because the original relationship between BPL and MPL statistics, as found in Altinok's data, does not assume, firstly, a normal distribution and, secondly, a similar standard deviation across all countries.

Figure 16: Baseline BPL-based patterns by world region (from tool)



The next two graphs depict the global distribution for 2015, and three future years: the target year 2030, and beyond that 2050 and 2070. Percentages refer to the percentage of the global child population to the right of the proficiency benchmark. Here the tool is set to use the MPL. **Figure 17** assumes a fast rate of progress, specifically the 'speed limit' described earlier. **Figure 18**, on the other hand, uses an optimistic business-as-usual improvement assumption based on trends seen in the international assessments in recent years. In this assumption, countries such as Morocco and South Africa are assumed to progress by 0.04 standard deviations a year, while Russia advances at 0.01 standard deviations a year. The label 'optimistic BAU' reflects the fact that this is the actual average trend seen in the historical statistics, but those statistics represent a group of relatively ambitious countries – this is why they choose to participate in an international assessment. The 'fast' scenario results in the percentage proficient value moving from 57% to 75% between 2015 and 2030, while the 'optimistic BAU' scenario sees an improvement from 57% to 65% over the same period. The latter scenario translates into a 0.5 percentage point improvement per year.

Figure 17: Four MPL-based distributions (fast)

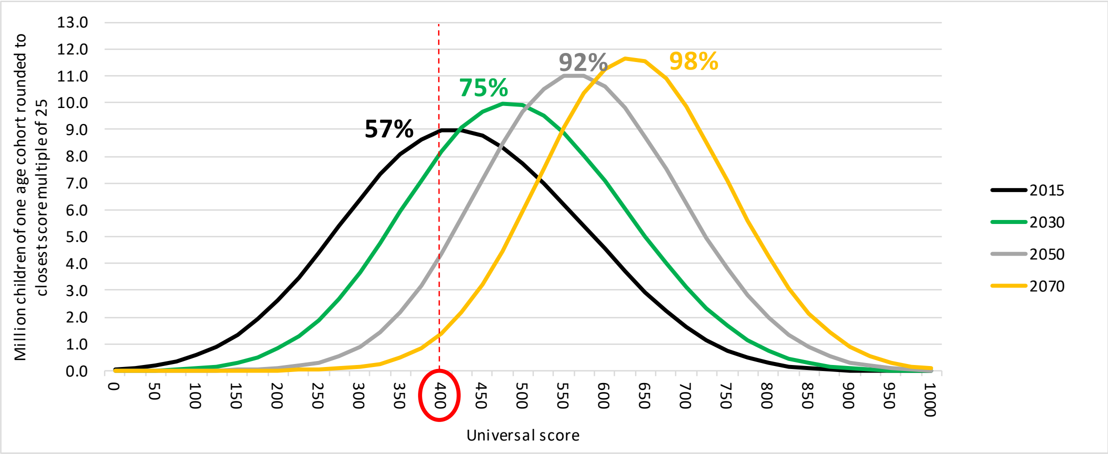
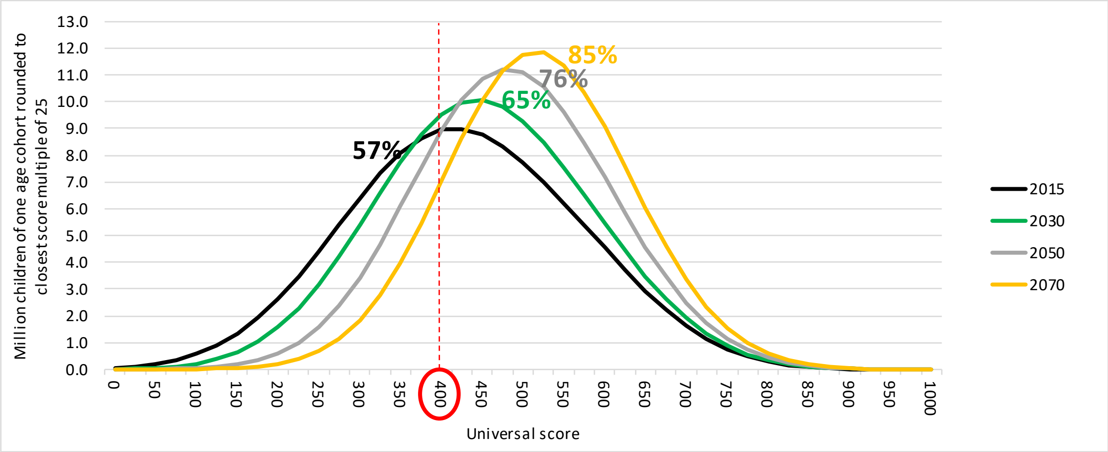


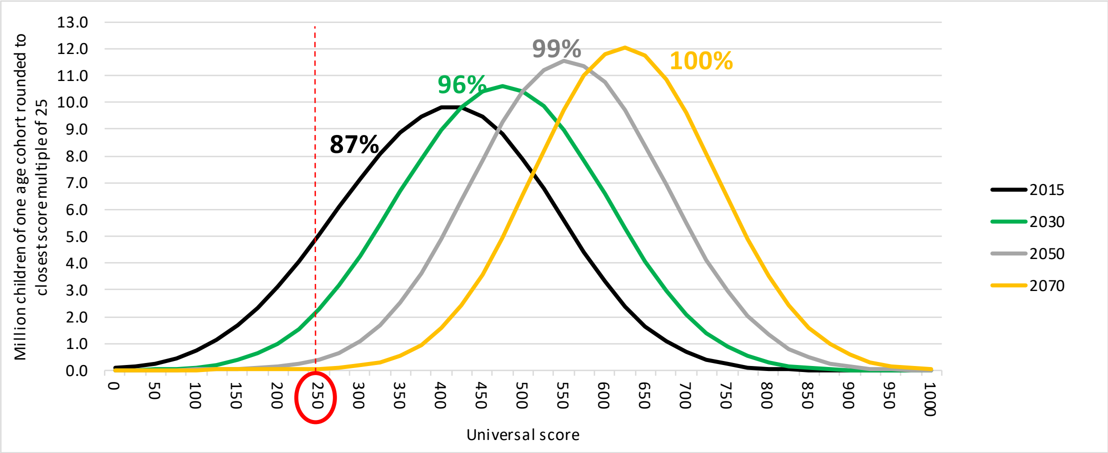
Figure 18: Four MPL-based distributions (optimistic BAU)



If differentiated standard deviations are used, specifically a standard deviation of 120 for a score of 250, and 72 for a score of 550 (see discussion above), future global percentage proficient statistics change a bit, as one would expect. Applying the differentiation to the projections seen in Figure 18 results in somewhat higher future values of 67% (2030), 79% (2050) and 88% (2070).

Figure 19 represents the BPL-based trend, and assumes everyone reaches the 'speed limit'. This is clearly an unrealistic assumption and, as discussed previously, BPL statistics reflect an interpretation of the learning challenge which denies the gravity of the problem and is not in line with the general discourse on learning outcomes. For the reason, BPL-based statistics were not replicated for dimensions other than lower primary reading.

Figure 19: Four BPL-based distributions (fast)



10 Margins of error in the baseline and projections

The fact that national proficiency statistics derived from datasets such as TIMSS are sample-based, means there are margins of error around these statistics. The margins are anywhere between 1.0 and 7.0

percentage points, half of this being on either side of the proficiency statistic. Such margins are implied by the standard errors published in the official TIMSS reports[48]. Those reports, and all the discussion presented here, assume a 95% confidence level. Developing countries tend to display larger margins of error. To illustrate, the average was a confidence interval of 5.7 percentage points across nine developing countries participating in TIMSS Grade 4 mathematics 2015: Bahrain, Chile, Indonesia, Iran, Jordan, Kuwait, Morocco, Saudi Arabia, and South Africa. Given that TIMSS is designed largely for high-income countries, such countries display much narrower confidence intervals, for instance 1.6 for Finland and 1.2 for Singapore.

One might expect developing country confidence intervals to be narrower in a programme designed specifically for these countries, but this is not necessarily the case. An analysis of SACMEQ 2007 reading data produces confidence intervals of between 3 and 10 percentage points, the mean across the 15 countries being 7 SACMEQ points[49]. It appears as if confidence intervals around proficiency statistics in SACMEQ are at least as large as those for developing countries in TIMSS.

Altinok's linking (or harmonisation) process widened many confidence intervals. An examination of Altinok's lower primary proficiency statistics points to a mean confidence interval across country and subject combinations of 5.0, with the 10th and 90th percentiles being 1.2 and 10.9 percentage points.

To calculate the confidence interval of a percentage proficient statistic for a world region (such as those in Table 9) the equation appearing below, which calculates a pooled standard error se_s, is needed. The version of the equation is the Satterthwaite approximation (hence the subscript *s*), which allows for different standard errors in different underlying samples[50]. One can illustrate the method through reference to the region Europe and Northern America, with its 50 countries. In general, standard errors for proficiency statistics in these countries are low, around 0.6 standard errors, producing confidence intervals of around 2.5 percentage points. To simplify the calculation, all values *s* can be assumed to be 0.6. Variable *n* represents the sample size for each country. Though the number of students in a testing programme such as TIMSS is typically about 4,000 per country, clustering by school means that the 'effective sample' is 400[51]. This effective sample is what TIMSS (and PIRLS and other programmes) aim for, and thus 400 should be used for *n*. The value *k* is 50 as there are 50 countries. Using these inputs, one would arrive at an se_s of 0.2, or a confidence interval of around 0.8. This represents less uncertainty than the uncertainty seen for individual countries. This confirms what one may intuitively expect: pooling results from various samples measuring the same thing results in less uncertainty in the aggregate statistics. Thus, at the regional level, a statistic such as 95% of children being proficient, in Table 9, would carry a confidence interval of 94.6% to 95.4%. This is a tiny interval, and movements within this band would be practically invisible in graphs such as Figure 14.

$$se_s = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \cdots + \frac{S_k^2}{n_k}} \qquad (5)$$

---

48 Mullis *et al*, 2016: Exhibit 2.2.
49 A proficiency level of 563 was used, which corresponds to SACMEQ 'Level 6', which in turn is roughly what Altinok (2017: 62) would consider the equivalent of the TIMSS 400 benchmark.
50 A simple explanation of this equation is provided at https://wolfweb.unr.edu/~ldyer/classes/396/PSE.pdf (title 'Pooled Standard Error versus the Satterthwaite Approximation').
51 Foy and Joncas, 2004: 114.

If one were dealing with a world region with larger standard errors around the proficiency statistics, what would the regional confidence interval be? If one takes the Latin America and the Caribbean region, assumes 48 countries, and a typical national standard error of 1.1 (this is the actual average across 16 countries for lower primary reading in this region in the Altinok data), se_S becomes 0.4, meaning a confidence interval of about 1.6 percentage points. Such an interval is less difficult to ignore.

What happens in the case of sub-Saharan Africa (SSA), a region where about 20 countries have no data at the primary level, and proficiency statistics must be imputed from, for instance, poverty data? A part of the answer lies in the standard errors produced when 19 of the countries in this region saw their lower primary reading statistics imputed, using income per capita, as explained in section 7.4. The mean standard error across the 19 countries was high, at 2.7, implying a confidence interval of around 10 percentage points in each country. This is just the uncertainty arising from the number of countries used to predict the relationship between income and proficiency, and the fact that one is predicting using income only, and no other variable. To these 10 percentage points, one should also add the uncertainty relating to the fact that proficiency statistics for all countries are based on samples of students. And this ignores uncertainty around whether countries falling outside testing systems such as PASEC and SACMEQ are especially weak performers, simply because there is a low level of awareness of educational quality. To roughly estimate the SSA regional standard error, the above formula was run with 30 countries carrying a standard error of 1.3 (an actual mean obtained from Altinok's upper primary data), and 20 with a standard error of 3.0 (the 2.7 value rounded up). The result is a standard error of 0.8 for the SSA region, or a confidence interval of 3 percentage points. This suggests we can be 95% certain that between 14.5% and 17.5% of children in SSA are proficient in the 2015 baseline – this is the confidence interval around the 16% mean reported in Table 9.

What about at the global level? What is the confidence interval around the 57% baseline figure seen in Figure 18? If one assumes 94 countries with income-imputed proficiency statistics displaying a standard error of 3.0, 50 generally rich countries displaying a standard error of 0.6, and the remaining countries, largely developing, with a standard error of 1.2, the resultant se_S is high, at 1.6, implying a confidence interval of 6 percentage points. However, this calculation ignores the fact that the 94 countries with income-imputed statistics tend to be small countries, and represent just 21% of the child population. It also ignores the fact that the 50 countries of Europe and Northern America represent just 10% of the world's children. If one adjusts the number of countries to reflect population, while retaining the total *k,* the result is a somewhat lower confidence interval of 5 percentage points. The global figure of 57% thus carries a confidence interval of more or less 54.5% to 59.5%.

Turning to margins of error around the projections, there is of course uncertainty associated with one's degree of optimism around the world's ability to improve beyond BAU trends. But uncertainties around what the actual past trends have been are also important. As discussed in the accompanying report, there is a risk that without proper examination of the quality of the available statistics, incorrect conclusions around past trends are easily arrived at.

The aim of this section has not been to provide an extensive and precise account of how the various processes of student sampling, across-programme linking, imputation of values using just Altinok's dataset, and imputations drawing from per capita income, have impacted on margins of error. It is very unlikely that a more detailed account would arrive at findings which are substantially different to the somewhat crude estimates presented above.

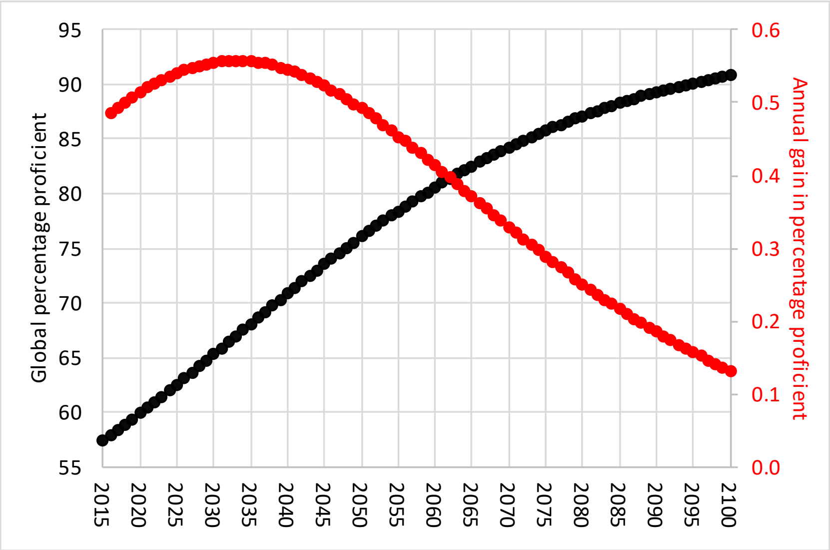
11 Extracting key information from the analysis and the tool

11.1 Changing rates of global change

As discussed in section 4, the fact that learning outcomes are distributed more or less along a normal curve, implies that a country starting from a low point of departure can initially expect relatively fast improvements, or improvements where an increasing number of children become proficient each year. After a turning point, specifically after the country's mean has surpassed the proficiency threshold, progress slows down. The number of children joining the 'stock' of proficient children declines each year.

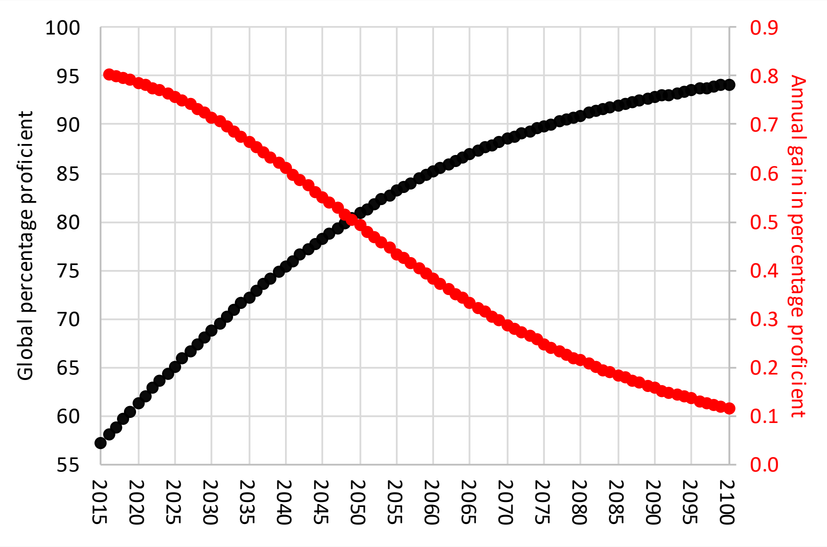
The baseline distributions seen in the above graphs, for instance Figure 18, suggest that the world has only just passed the proficiency threshold of 400. However, when the 'optimistic BAU' scenario reflected in Figure 18 is used, the annual changes reflected in **Figure 20** are found. Figure 20 illustrates an increasing rate of improvement up to around 2035, in the sense that the annual gain in the global percentage proficient statistic gets larger up to this point, after which annual gains becomes smaller (see the red curve). This may seem odd given that the world mean appears to have passed the 400 threshold (Figure 18). **Figure 21** explains the apparent contradiction. Figure 21 is like Figure 20, except the baseline population is kept unchanged for all future years. In Figure 21, the gain in the percentage proficient declines in all years. This points to an important dynamic. Relatively high population growth in a region whose mean is clearly to the left of the 400 threshold currently, namely sub-Saharan Africa, means that increasing gains in the global indicator are likely, *if* the improvements assumed by the 'optimistic BAU' scenario are achieved.

Figure 20: Rate of global change up to 2100



The red curve should be read against the right-hand vertical axis.

Figure 21: Rate of global change up to 2100 with no population change



The very large influence of sub-Saharan Africa on any global trends based on population is apparent in **Figure 22** below. The UNPD medium variant projections described above envisage the percentage of children accounted for by sub-Saharan Africa increasing from 23% in 2020 to 46% in 2100. By 2030, this figure is expected to be 27%.

Figure 22: Population growth by region to 2100

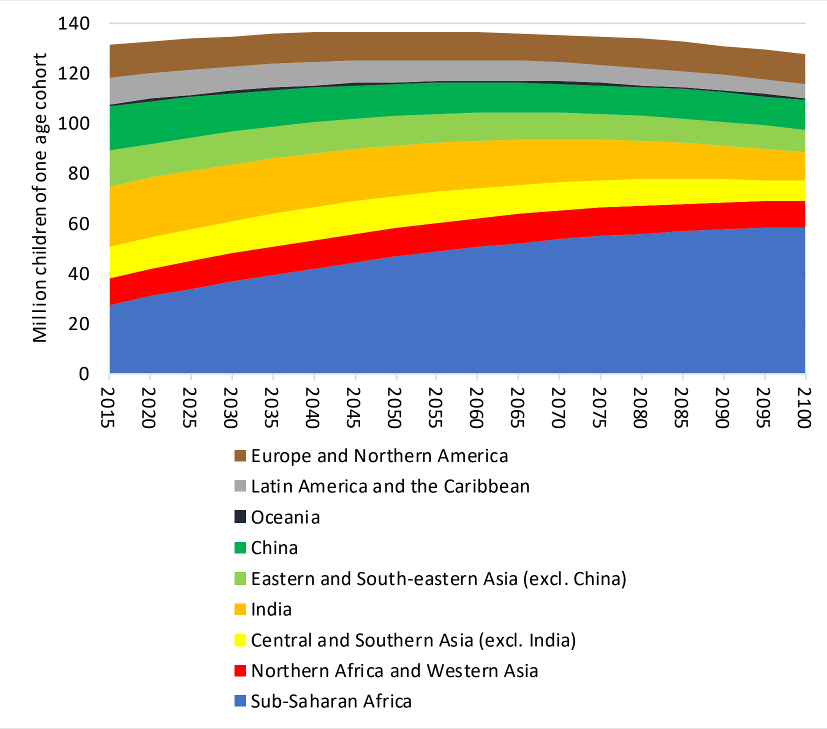


Figure 23, which is based on the UNPD medium variant projections, serves as a reminder that there are exceptions. Within sub-Saharan Africa, several Southern African countries do not experience rapid population growth. Outside this region, countries such as Iraq and Papua New Guinea are expected to see their child populations increase by over 20% in the 2020 to 2040 period.

Figure 23 Age 5 to 9 population growth by country 2020 to 2040

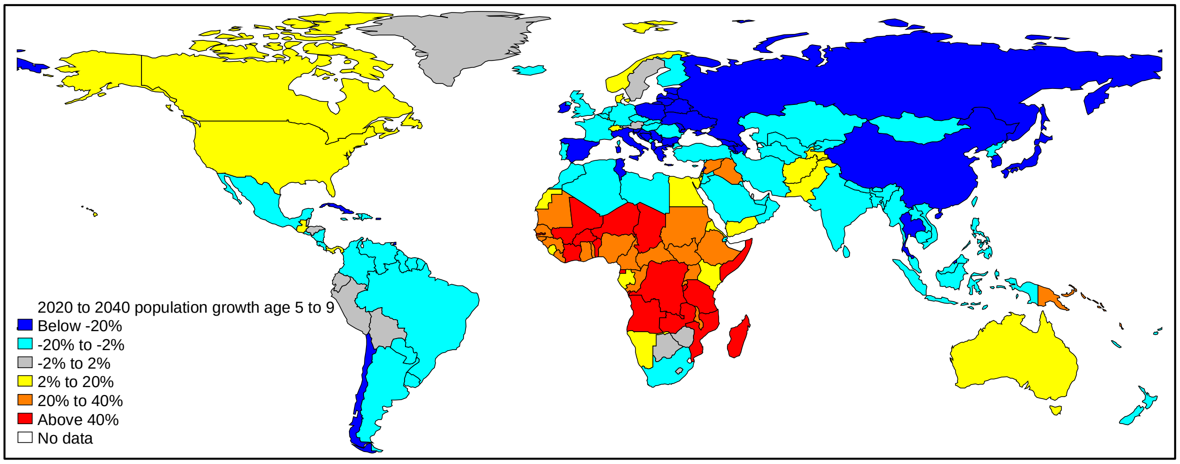
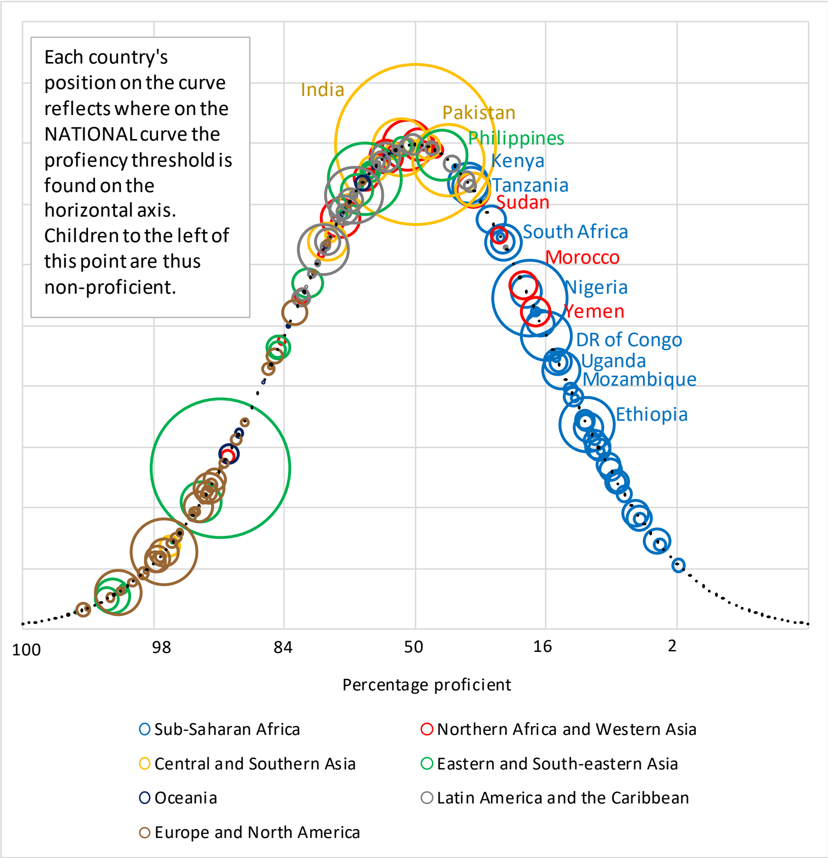


Figure 24 below offers one way of visualising how countries progress and how rates of improvement change. The curve is a normal one. Countries are situated on this curve in accordance with where the 400 benchmark is found in the baseline year. Lower primary reading figures are reflected. For example, Ethiopia sees the 400 benchmark on the far right of its distribution. Only 13% of children are above the benchmark in the baseline year. As the country improves, it will move leftwards, as more children move to the right of the 400 score mark. The upward movement along the curve reflects the fact that each year an increasing number of children will enter the ranks of the proficient. This will continue until half of Ethiopia's children surpass the 400 mark, after which annual gains will begin to shrink.

Figure 24: Countries with a large potential for contributing to global change

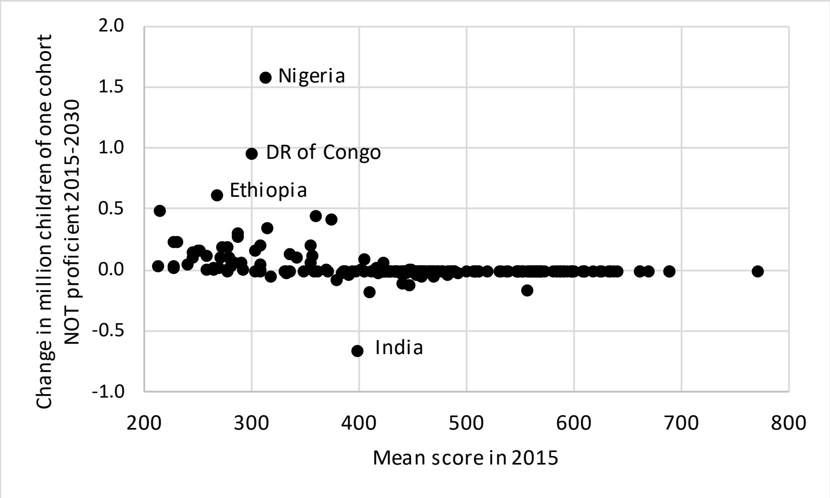


The above graph thus helps to visualise the fact that although sub-Saharan African countries are in general behind with respect to learning proficiency, with the right action they can expect exceptionally steep improvements.

11.2 The risk of a demographically driven move backwards

The fact that the tool allows for the combined effect of educational progress and demographic change to be analysed is important. One important risk is easily illustrated by the tool. If one assumes, probably unrealistically, that all countries experience zero progress in their mean score, and hence in their percentage proficient, then the global percentage of proficient children would actually *decline*, from 57% to 53% over the period 2015 to 2030 simply due to the changing demographic weight of countries. **Figure 25** draws from the tool where zero country-level progress is assumed, the selected indicator being reading at the lower primary level. In this scenario, countries such as Nigeria, Democratic Republic of Congo and Ethiopia could add, in absolute terms, many children to the stock of non-proficient children, unless of course sufficient attention went towards strengthening schooling in these countries.

Figure 25: Possible increases in non-proficient children (early primary reading)



One can use the tool to explore scenarios in which the global percentage proficient (for lower primary reading) would remain at 57% up to 2030. One way of achieving this would be to have an annual improvement of 0.01 standard deviations across all countries in the world. In other words, this level of improvement would be necessary simply for the world to keep up with demographic trends.

11.3 Ambitious yet evidence-based targets per region

A key purpose the current report and the accompanying Excel tool is to provide both global and regional disaggregated baselines and target figures for educational proficiency. As discussed previously, patterns seen at, say, the regional level, provide an indication of what national education authorities should aim for, and how they should understand progress, though the national targets available in the tool are mostly based on baseline data which are insufficiently interrogated for them to be used 'as is'. **Table 20** provides 2030 targets, or projected proficiency levels, using the optimistic BAU assumptions explained in relation to Figure 18 above. The 2015 baseline corresponding to Table 20 is what is shown in the earlier Table 11. Clearly, across most regions and dimensions, targets of 100% proficiency do not emerge as even remotely possible. In sub-Saharan Africa, it seems especially important to manage expectations. Reaching a level of reading proficiency at the lower primary level of 33% may not seem good, but the evidence suggests this would be a remarkable step forward for the region. While it is unlikely that any political leader would want to put forward a target as low as this, it is important for monitoring staff in national authorities to be fully aware of the actual dynamics, and to be ready to explain why even apparently modest gains were important development achievements.

Table 20: Ambitious yet evidence-based 2030 targets across all dimensions

	Lower primary		End of primary		End of lower secondary	
	Read.	Math.	Read.	Math.	Read.	Math.
Sub-Saharan Africa	33	44	29	38	24	29
Northern Africa and Western Asia	65	64	65	65	51	57
Central and Southern Asia	66	68	66	68	40	42
Eastern and South-eastern Asia	88	84	88	84	73	76
Oceania	87	86	87	86	77	80
Latin America and the Caribbean	78	75	84	68	66	60
Europe and Northern America	97	95	97	95	87	91
World	65	68	65	65	50	53

11.4 Implications for cross-national monitoring efforts

A few things emerge, or are confirmed, in the current report around how to take the monitoring work of the UIS, and other such organisations, forward.

When a serious attempt is made to project learning outcomes, and set targets, a key hurdle one still encounters is insufficient knowledge about how fast one can reasonably expect these outcomes to improve, and what measures one should use. With a stronger focus on improvement, in particular historical improvements in specific countries, future research should be able to address much of this knowledge gap. A part of the task is to get more time series data in more countries, especially developing countries, and to pay considerable attention to the quality of the data.

The report confirms that what is now widely understood as the 'minimum proficiency level', or MPL, which equals 400 TIMSS points is a useful threshold for monitoring progress with respect to indicator 4.1.1. The existing narrative that around one-half of the world's children and adolescents do not reach a minimum level of proficiency is supported by the 400 TIMSS points MPL. The indicator values for sub-Saharan Africa, such as only 16% of children in this region at the lower primary level reach the MPL in reading, may seem daunting, yet they are useful. It is worth bearing in mind that the monitoring of learning outcomes does not suffer from the non-normality problem seen in the monitoring of poverty. If 84% of people are poor, that information could be deceptive as the distribution of, say, income is non-normal, meaning important questions arise around what part of the distribution to focus on. However, if 84% of children cannot read properly, the evidence suggests the distribution of the non-proficient children will follow a normal curve. The justification for focussing on everyone when it comes to educational proficiency is stronger than when it comes to poverty. Schooling systems are systems, and they tend to improve in relatively predictable ways. Of course, there are some important questions around sub-categories of children among the non-proficient. For example, whether either boys or girls are falling behind has important policy implications.

The question of whether expectations at different levels of the schooling system should be set in such a way that proficiency percentages are in general higher at, say, the lower primary level than the lower secondary level, is an unresolved yet important question. The data used for the tool produce lower levels of proficiency at the secondary level than at the primary level, a pattern many education planners would argue is appropriate and produces the right messages. However, this pattern comes about by accident. It was not built into the monitoring procedures described in the current report. The question is whether it should be. Lower proficiency levels at the secondary level produce the right messages if there is an understanding that improving proficiency at the secondary level means, above all, improving proficiency at the primary level, and if this promotes work on enhancing secondary education's role in preparing youths for the world of work and active citizenry. However, lower levels of proficiency at the secondary level should generally not detract from the importance of improving the foundations for learning at the primary level. In short, differences in proficiency statistics across the levels of the schooling system need to be considered in the design of monitoring systems, and the meaning of these differences needs to be communicated clearly.

The report has emphasised the importance of taking into account population change across the world. It is possible for no global improvement to be occurring with respect to proficiency, while some countries experience zero improvements, and others *do* experience improvements. If statistics are improving in, say, sub-Saharan Africa in a context where global statistics are not, this is not all bad. Global monitoring needs to be clear about how educational improvement and demographic trends are separately influencing the global picture.

The finding that not just national distributions of test scores, but also those of world regions, and even the world, follow a normal distribution, is important. This facilitates planning and monitoring at the global and regional levels.

Finally, and importantly, apparent completeness in the global statistical tables should not detract from the fact that much of the data on educational quality, and in particular on national quality trends, is weak. As Dan Inbar, in a UNESCO planning guide, pointed out: "Comprehensive knowledge, organized into a coherent framework, gives the appearance of better control, or reduced uncertainty, and a decrease of risks"[52]. The fact that there are serious issues with the underlying national statistics should not be forgotten. There should be a continual push for better data, and available data must be utilised and scrutinised better. Uncertainties in global and regional trends should be made explicit.

11.5 Implications for national monitoring efforts

This analysis confirms two particularly important points for national planners, and those providing technical support for national planners.

First, targets that are feasible and believable, in terms of the best improvements seen historically at the country level, need to be far more modest than what is implied by the SDG target on learning proficiency. This creates difficult institutional and political problems, which are not easy to resolve. It is very likely that politicians will reject modest, and realistic targets, as they are extremely difficult to 'sell' politically, despite what the technocrats say. This in turn creates further problems. National monitoring systems may not be set up to detect small improvements, meaning that if these occur, they are not appreciated as they should be. Non-attainment of high aspirational targets in fact sets politicians up for failure, meaning politicians may in the future try to divert attention away from educational quality to non-education issues or education issues that are unrelated to quality. The best way of dealing with all these risks is probably for the technocrats

[52] Inbar, 1996: .91

to develop themselves technically, fine-tune the monitoring systems, persist in focussing on modest gains, and explain convincingly why, when these occur, this is a vital sign of national development.

Second, it is clear that the normality of test score distributions needs to be taken into account in conversations about realistic targets. Where the minimum proficiency level is found in the national curve has important implications for the magnitude of the future improvements one can expect.

References

Altinok, N. (2017). *Mind the gap: Proposal for a standardised measure for SDG 4 – Education 2030 agenda.* Montreal: UIS.

Birchler, K. & Michaelowa, K. (2016). Making aid work for education in developing countries: An analysis of aid effectiveness for primary education coverage and quality. *International Journal of Educational Development,* 48: 37-52.

Botswana (2015). *Education & training sector strategic plan (ETSSP2015-2020).* Gaborone.

Bruns, B., Evans, D. & Luque, J. (2012). *Achieving world class education in Brazil: The next agenda*. Washington: World Bank.

China: Ministry of Education (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> [Accessed: August 2018].

Crouch, L. & Gustafsson, M. (2018). *Worldwide inequality and poverty in cognitive results: Cross-sectional evidence and time-based trends*. Oxford: Research on Improving Systems of Education.

Foy, P. & Joncas, M. (2004). *Chapter 5: TIMSS 2003 sampling design.* Chestnut Hill: IEA.

Gustafsson, M. (2011). *Floor effects and the comparability of developing country student test scores*. Paris: UNESCO. [Forthcoming.]

Hanushek, E.A. & Woessman, L. (2007). *The role of school improvement in economic development*. Washington: National Bureau of Economic Research.

Ho, D.H. & Yu, C.C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, 75(3): 365-388.

IMF (2000). *2000 a better world for all: Progress towards the international development goals.* Washington.

Inbar, D.E. (1996). *Planning for innovation in education*. Paris: IIEP.

India: National Council of Education Research and Training (2018). *National Achievement Survey.* Delhi. Available from: <https://mhrd.gov.in/sites/upload_files/mhrd/files/upload_document/State-Report-Card-21st-march.pdf> [Accessed: August 2018].

Intergovernmental Panel on Climate Change (2006). *2006 IPCC guidelines for national greenhouse gas inventories*. Hayama: IPCC.

Lakner, C., Mahler, D.G., Negre, M. & Prydz, E.B. (2019). *How much does reducing inequality matter for global poverty?* Washington: World Bank.

Mingat, A., Rakotomalala, R. & Tan, J.-P. (2002). *Financing Education for All by 2015: Simulations for 33 African countries*. Washington: World Bank.

Mullis, I.V.S., Martin, M.O., Foy, P. & Drucker, K.T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill: Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill: IEA.

Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Chestnut Hill: Boston College.

OECD (2017). *PISA 2015 results: Students' financial literacy*. Paris.

Stukel, D.M. & Feroz-Zada, Y. (2010). *Measuring educational participation: Analysis of data quality and methodology based on ten studies*. Montreal: UNESCO Institute for Statistics.

Taylor, S. & Spaull, N. (2015). Measuring access to learning over a period of increased access to schooling: The case of Southern and Eastern Africa since 2000. *International Journal of Educational Development*.

UIS (2017a). *More than one-half of children and adolescents are not learning worldwide*. Montreal.

UIS (2017b). *Counting the number of children not learning: Methodology for a global composite indicator for education*. Montreal.
UIS (2018). *GAML Brief 6: The anchoring method for Indicator 4.1.1 reporting*. Montreal.
UIS (2018). *Costs and benefits of different approaches to measuring the learning proficiency of students (SDG Indicator 4.1.1)*. Montreal.
UIS (2019). *How fast can levels of proficiency improve? Examining historical trends to inform SDG 4.1.1 scenarios*. Montreal.
UNESCO (2002). *Education for All global monitoring report 2002: Is the world on track?* Paris.
UNESCO (2015). *EFA Global Monitoring Report 2015: Education for all 2000-2015: Achievements and challenges*. Paris.
UNESCO (2016). *Getting climate-ready: A guide for schools on climate action*. Paris.
UNESCO (2017). *Global Education Monitoring Report 2017/18: Accountability in education: Meeting our commitments*. Paris.
UNESCO (2018). *Global Education Monitoring Report 2019: Migration, Displacement and Education – Building Bridges, not Walls. Paris.* Paris.
United Nations (2000). *United Nations Millennium Declaration*. New York.
United Nations (2015). *The Millenium Development Goals Report.* New York.
United Nations (2017a). *Guidelines to support country reporting on the Sustainable Development Goals.* New York.
United Nations (2017b). *Revised list of global Sustainable Development Goal indicators*. New York.
United Nations: Department of Economic and Social Affairs (2019). *World Population Prospects 2019: Data booklet.* New York.
Walker, M. (2011). *PISA 2009 plus results*. Camberwell: Australian Council for Educational Research.
Wößmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics,* 13(2): 143-169.