United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

SUSTAINABLE
DEVELOPMENT
GOALS

Information Paper No. 62
December 2019

How Fast Can Levels of Proficiency Improve?

Examining Historical Trends to Inform SDG 4.1.1 Scenarios

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). S

Published in 2019 by:

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

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

Ref: UIS/2019/LO/IP62

Summary

This report summarises the existing analysis, and adds new analysis, on historical trends with respect to learning outcomes in schooling systems around the world, with a view to informing projections for Sustainable Development Goal (SDG) indicator 4.1.1, which deals with the percentage of children who are proficient in reading or mathematics. Trends with regard to two statistics receive attention: mean scores of countries, and the percentage of students reaching specific levels of proficiency. Though SDG 4.1.1 concentrates on the latter, both statistics are important for gauging and modelling qualitative change.

The report begins with a review of existing work of relevance to the topic. Analyses that have pointed to flaws in the existing international testing programmes, flaws which can produce inaccurate trends, serve as a reminder that care must be taken when interpreting the data from these programmes. Governments of both developed and developing countries release targets for improvements which tend to be over-ambitious, and not based on evidence of what is possible. This underlines the importance of the type of analysis presented here. Definitions, including those relating to SDG 4.1.1, can be ambiguous. Clearly, definitions must be clarified before trends are calculated. Equating scores across different international testing programmes is one way of enlarging the sample of countries used when historical trends are determined. Careful use of standard deviations as a metric of improvement in the mean score can serve the same purpose. Where in the distribution of scores minimum proficiency levels are set influences in various ways trends with regard to the percentage of proficient children.

Turning to the central concern of the current report, not much analysis of historical trends, in the sense of either the global business-as-usual trend or best possible trends, or 'speed limits', has occurred. This represents a serious gap for planners, who need to set targets and should draw from evidence of past trends. The analysis that exists has focussed on improvements in the mean score. Annual improvements, at a country level, of between 0.01 and 0.06 standard deviations, can be found in the existing analyses. Importantly, country-wide improvements seen in the data should not be confused with the higher improvements seen in much of the experimental research focussing on samples of schools. For various reasons, it appears virtually impossible to replicate the magnitude of improvements seen in the latter, across whole countries.

The new analysis presented in the report draws from the data of the PIRLS, PISA and LLECE programmes, and focusses on both the primary and secondary levels, and both of the subjects reading and mathematics. It is concluded that the business-as-usual (BAU) trend in recent years points to larger gains for educationally less developed countries. It is easier to improve off a lower base. This would be in line with the notion that there are diminishing returns to educational improvement efforts as countries develop, and that 'natural ceilings' for cognitive skills exist. All three programmes point to similar business-as-usual and 'speed limit' values. The proportion of countries displaying positive slopes in the mean, or improvement over time, comes to between two-thirds (PIRLS lower primary reading) and almost 90% (LLECE upper primary mathematics). Ongoing improvement thus appears to be the norm rather than the exception. All three programmes point rather consistently to a BAU improvement trend of around 0.04 standard deviations a year for the least

developed countries, and between zero and 0.01 for developed countries. Moreover, they suggest a 'speed limit' of around 0.08 standard deviations a year for the least developed countries. There are a few developing countries which have exceeded this, with for instance Qatar, based on three years of PISA data, reaching almost 0.09. However, improvements in excess of this 0.08 limit are extremely rare, meaning it seems realistic to set, as a very ambitious though not impossible target, 0.08 standard deviations a year.

Turning to the percentage of children who are proficient, among the worst performing PIRLS countries annual improvements of 2.0 percentage points are the trend, though for the best performing countries, the norm is zero. The latter is not surprising as well performing countries are left with very few non-proficient children. These figures can be considered a basis for business-as-usual scenarios focussing on the percentage proficient.

The report concludes that the findings presented here constitute a good basis for determining realistic and empirically informed, though not necessarily politically palatable, SDG 4.1.1 targets. The analysis could be replicated, using further data sources. However, there are other, more serious, research gaps. In particular, there is very little multi-disciplinary research aimed at understanding what policy and socio-economic factors explain how certain countries, such as Indonesia and Trinidad and Tobago, have achieved exceptional educational improvements over time.

Contents

Summary **3**
1. Introduction **6**
2. Existing work of relevance to understanding historical trends **7**
2.1 Analysis into the reliability of existing test score trends 7
2.2 How countries set targets 8
2.3 SDG 4.1.1 definitions and intentions 9
2.4 Equating the scores of different testing programmes 10
2.5 Minimum proficiency cut-scores and calculating percentage proficient 12
2.6 Improvements in terms of fractions of standard deviations 13
2.7 Existing work on business-as-usual rates and 'speed limits' 13
3. New analysis of change in terms of means and percentage proficient **16**
3.1 Evidence from PIRLS 16
3.2 Evidence from PISA 22
3.3 Evidence from LLECE 23
4. Concluding remarks **24**

1. Introduction

The Sustainable Development Goals (SDGs), agreed upon within a resolution adopted in the United Nations General Assembly on 25 September 2015, include as Goal 4.1 the following[1]:

4.1 By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes

In 2017, the following indicators were agreed on for the above goal[2]:

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

Indicator 4.1.1 is striking because compared to the earlier Millenium Development Goals (MDGs) governing the global development agenda between 2000 and 2015, this SDG indicator enhances enormously the focus on learning, as opposed to mere attendance. This reflects the fact that today, to a far greater degree than in around 2000, the economic and development literature puts human skills at the centre of human progress and survival, and as a means for reducing social inequality[3].

The word 'all' in the goal implies a target for the indicator of 100% by 2030. The SDG targets in general, and in particular those of SDG 4.1.1, are set politically, and have not really been informed by scientific projections of what is possible. Aspirational targets are a reality in global and national politics, yet credible projections are necessary to inform, at the very least, reporting and conclusions around whether systems are truly succeeding or failing when they do not reach targets.

The current report is intended to facilitate credible projections by summarising existing analyses, and presenting new analysis of historical trends seen in international testing programmes. The assumption is that credible projections need to be informed by rates of improvement experienced by whole schooling systems in the past. For reasons explained below, trends with respect to both mean scores and the percentage of students who are proficient, receive attention.

Section 2 reviews the existing literature. First, critiques of the reliability of the trends emerging from the international programmes are reviewed (2.1). The question of how education authorities currently set test score or proficiency targets for a whole system is addressed (2.2). This is important to understand, as the report is concerned with improving this process. Definitional issues relating to SDG indicator 4.1.1 are then discussed (2.3). One way of dealing with the reality that different international programmes use different metrics, is to recalibrate existing measures to one common metric. How this has been attempted, and the

[1] United Nations, 2015.
[2] United Nations, 2017.
[3] OECD, 2010; UNESCO, 2016: 24.

limitations of the results, is discussed (2.4). Another way of imposing a 'common language' across testing systems is to talk of progress in mean scores in terms of standard deviations. But this comes with its own comparability caveats (2.6). Gauging progress in terms of the percentage of students reaching minimum proficiency levels obviously requires determining what these levels are. How this determination has occurred in different international programmes is discussed (2.5). The review then proceeds to the central question of how previous analysts have understood national and global rates of progress (2.7). It is speculated why there has not been more work in an area of such obvious importance for education planners. The importance of distinguishing between system-wide rates of progress and progress seen in samples of schools participating in research programmes is emphasised.

Section 3 presents new analysis on how fast the world and countries have progressed in recent history. The focus is two-pronged. On the one hand, what the business-as-usual rates of improvement have been are examined. On the other, the fastest believable improvements are identified as these can be considered 'speed limits' which planners should use. Two statistics receive attention: mean scores and the percentage of students who are proficient (or 'percentage proficient'). The first statistic is important in its own right, and has received considerable attention among analysts attempting to gauge historical progress. Many government planners focus on this statistic, and it makes sense to link it in some way to SDG 4.1.1. The second statistic is important insofar as this is precisely what SDG 4.1.1 focusses on.

Section 4 concludes, in part by pointing out what further work seems important on the road ahead.
The current report is accompanied by, and feeds into, a second report titled *Projecting attainment of SDG 4.1.1.*

2. Existing work of relevance to understanding historical trends

2.1 Analysis into the reliability of existing test score trends

Official national averages emerging from international and national assessment programmes are widely publicised and discussed. What has received less attention, are noteworthy critiques of the reliability of the apparent trends.

Jerrim (2013a) argues that due to changes in the way testing occurred, a downward trend in PISA[4] in the case of England was deceptive. There was probably no decline. At the same time, Jerrim (2013b) finds that PISA scores in general are reliable and an important basis for drawing conclusions about educational progress.

More importantly in terms of understanding improvements in developing countries, Brazil's PISA mathematics improvements in the 2000 to 2009 period, which have been put forward as exemplary for developing countries, have been questioned. Specifically, analysis by Klein (2011) and Carnoy *et al* (2015)

[4] Programme for International Student Assessment.

suggest that improvements probably came to around 3 PISA points a year, and not the 6 points a year suggested by the published figures. Klein suggests that the distortions affecting Brazil, which related largely to changes in the date of testing, also affected the PISA scores of other Latin American countries. These problems are acknowledged in the World Bank's 2018 *World Development Report*[5].

If trends with respect to average scores are unreliable, then trends with respect to the percentage proficient would likewise be problematic. Cases where trends appear unreliable should be carefully considered when 'speed limits', or the best feasible trajectories, are identified.

2.2 How countries set targets

If the aim is to arrive at methods to set feasible targets for SDG 4.1.1 for national authorities, one point of departure has to be how countries currently set these kinds of targets. Where countries do set targets in relation to percentage proficient, or average scores, they tend to be unrealistically high. This should not come as a surprise. In the absence of any technical guidance, and in an area as complex and politicised as improvements in educational quality, it is understandable that over-ambitious targets should be set. The fact that existing targets tend to be higher than they should be is what informs the use of the term 'speed limits' in this report. There are speeds of improvement which appear to be unattainable, based on what has happened in the past. This is not to say that new teaching and learning techniques, such as those using individualised instruction through computers, will not change the 'speed limits'. However, until such innovations are realised, and can be proven to work on a large scale, it is necessary to base expectations to a large degree on what has previously been achieved.

Examples of unrealistically high targets are not limited to developing countries. The US state of Hawaii, as part of the No Child Left Behind policy, envisaged in a 2003 plan that students in disadvantaged schools who are proficient in mathematics would increase from a baseline of 20% in 2001 to 100% in 2014[6], so by around six percentage points a year. In Ethiopia, a 2015 plan envisages that the percentage of Grade 2 pupils reaching a basic level of literacy would reach 70% by 2019, off a baseline of around 25% for 2014. This comes to nine percentage points a year. The same plan acknowledges that though an earlier plan had envisaged 75% of Grade 4 pupils reaching a minimum level of competency by 2012, what was realised by 2012 was just 25%[7]. The discussion in section 3 below will make it clear that the targets set in Hawaii and Ethiopia were clearly 'off the radar' as far as actual improvements seen in the recent past are concerned. The targets were overly ambitious, and would, if interpreted mechanistically, have made even excellent progress appear as a failure.

Some searching through UNESCO IIEP's Planipolis repository[8] of education plans reveals that most countries have not formally set targets for percentage proficient, or average scores. This is fortunate insofar as it

[5] World Bank, 2018: 93.
[6] Hawaii: Department of Education, 2003.
[7] Ethiopia: Federal Ministry of Education, 2015: 18, 40.
[8] http://planipolis.iiep.unesco.org.

means introducing new target-setting methods does not mean overturning previously agreed upon, and unambitiously high, targets, in the case of most countries.

2.3 SDG 4.1.1 definitions and intentions

Indicator 4.1.1, whose official description was given in section 1, is clear enough for overall policy discussion purposes, yet there are ambiguities one needs to grapple with on a technical level, when one calculates the proportion.

Critically, is the denominator 'children and young people' generally, in the sense a cohort of the population of the age commonly associated with one of the three education levels mentioned? Or is it only people *in* school at the levels mentioned? The *goal*, Goal 4.1 quoted above, suggests the former. However, if one looks just at the *indicator* description, it appears that the denominator is school students, excluding the parts of the population not in school. However, this is not totally clear. Much depends on the meaning of the word 'in'. Is it just the numerator, or both the numerator and denominator, which is limited to people *in* schools? A 2017 UNESCO Institute for Statistics metadata report employs the latter understanding of the denominator being just *students*[9]. This definitional matter is important if one considers that globally around 61 million children who should be in primary school are not (compared to total primary enrolments of 724 million)[10]. In the new analysis presented below, the current report does not take into consideration children not in school. However, one can presume there would be considerable political pressure to consider *all* children of an age cohort, whether in school or not, as the denominator. Doing this is probably not as difficult as the metadata report makes out. A fairly commonly used method is to assume that all non-enrolled children perform below any minimum standard[11].

There is another, less obvious, reason why ambiguities relating to the denominator are concerning. This second concern relates to grade repetition. In many developing countries, levels of grade repetition are high. To take a not too unusual example, in Burundi around 25% of Grade 6 students were repeating this grade in 2011 according to UNESCO data[12]. The logic behind grade repetition is that one is forced to repeat by the system because one has not achieved the required minimum level of performance. Let us assume that all students in Burundi *do* reach a minimum level of performance in Grade 6, eventually, though a third of them only succeed in doing so after spending two years is this grade. (If 25% of students are repeaters, this comes to a third of the remaining 75% representing first-time Grade 6 students. This means a third of all students repeat Grade 6.) In such a theoretical scenario, SDG 4.1.1 could display a value of 75% *even if 100% of children were achieving a minimum level of proficiency*. The remaining 25% would be students 'on hold', waiting to pass on their second attempt. This is of course a hypothetical and extremely unlikely scenario, and we can assume that many students in Burundi do *not* become minimally proficient, even after repeating. However, even in

[9] UNESCO Institute for Statistics, 2017a: 8.
[10] UNESCO, 2017: 320
[11] See for instance Hanushek and Woessman (2007).
[12] Indicator 'Percentage of repeaters in Grade 6 of primary education, both sexes (%)' off UIS.Stat at http://data.uis.unesco.org, accessed July 2018.

a more realistic scenario, SDG 4.1.1 statistics would almost certainly *under*-estimate, to some degree, the actual attainment of proficiency in a scenario where there is grade repetition.

Turning to the influence of *changes* in grade repetition on proficiency trends, the following table represents what could easily be the situation in many countries. Figures are for South Africa. The percentage of students in Grade 9 mathematics in TIMSS[13] who reached the TIMSS 'low international benchmark' increased from 24% to 34% between 2011 and 2015. This is gain of 10.0 percentage points. In South Africa, grade repetition has been declining, the average across all grades (not Grade 9 specifically) having declined over five years from 11.8% of students repeating their grade in a year, to 9.6%. If we assume a constant age cohort of 100.0, then students in a grade declined from 113.4 to 110.6. The number of students who are proficient would thus be 27.2 (24% of 113.4) and 37.6 (34% of 110.6) in the two years. Thus the percentage of an *age cohort* achieving proficiency would be 27.2% in 2011 and 37.6% in 2015. This represents a gain of 10.4 percentage points (37.6 minus 27.2), not the initially apparent 10.0 percentage points.

Table 1: Grade repetition and proficiency statistics

	2011	2015	Gain
Proficiency statistic	24	34	10.0
% repeating	11.8	9.6	
Size of an age cohort	100.0	100.0	
Size of the grade	113.4	110.6	
% of an age cohort proficient	27.2	37.6	10.4

While in this example taking repetition into account did not change the gain to a large extent, it is conceivable that in less typical situations, the effect of changes in grade repetition would be large.

The UNESCO data indicate that of 121 countries with data, 12 displayed an average percentage of repeaters value for Grade 6 and the years 2013 to 2018 exceeding 15%. Of the 121, 23 countries displayed an annual downward slope steeper than -0.55, meaning a more rapid decline in grade repetition than that seen in South Africa. These figures seem to confirm the importance of taking into account the effects of grade repetition in the global monitoring of proficiency trends. The argument has been made here in terms of the percentage proficient. A similar argument could be made for the effect grade repetition on the mean score.

2.4 Equating the scores of different testing programmes

Some sense is needed of the equivalence of different testing programmes. Global monitoring requires this. If one is looking at the percentage of students passing a minimum level of performance in countries A and B, or regions X and Y, one would want these minimum levels to be at least roughly equivalent. If they are not absolutely equivalent, the policy implications are not too serious. What is more important from a development and SDG perspective is that within each country measurement is consistent over time, so that

[13] Trends in International Mathematics and Science Study.

progress can be properly gauged. Achieving very high levels of comparability *across* countries would be a costly, complex and politically difficult task. Fortunately, achieving consistency in the statistics *within* a country, while not simple, is a less difficult task. Yet across-country comparability should be pursued as far as possible, and much work in this regard has already occurred. Achieving across-programme comparability largely means equating the scores of international testing programmes, such as SACMEQ[14] and PISA, with one serving as a standard. However, it is arguably also about equating at least some national assessment programmes to a global scale. Some national programmes, such as India's emerging National Achievement Survey, represent more students than some of the regional programmes.

So what work in equating different assessment programmes has occurred? In this section, attempts to equate scores are discussed. In the next section, the determination of global proficiency thresholds, or cut scores, a task which is a somewhat separate one, is discussed.

Sandefur (2016: 13), in attempting to equate SACMEQ Grade 6 and TIMSS Grade 8 mathematics scores, explores three different linking methods: equipercentile linking and two linking approaches making use of the fact that some TIMSS test items were included in SACMEQ. For example, a mean of 600 in SACMEQ translates to around 400 in TIMSS using the equipercentile approach. The fact that different grades are covered by the two programmes does not really matter, in particular if one makes the assumption that qualitative progress between grades 6 and 8 in each country is not so different across countries that rankings change. One can think of, say, the Grade 6 SACMEQ average score for Swaziland (around 518) converted to the TIMSS Grade 8 scale (around 350) as what Swaziland's students *would* achieve when in Grade 8. The average scores of fourteen African countries from SACMEQ are expressed in terms of both TIMSS Grade 8 and TIMSS Grade 4 scales.

Altinok *et al* (2018: 37) use just 'presmoothed equipercentile' linking to produce ten two-way translations between international assessment programmes. To illustrate, the translation between LLECE[15] 1997 Grade 6 and TIMSS 1995 Grade 8, in both cases mathematics, specifies that a mean and standard deviation of 489 and 71 in LLECE translate to a mean of 344 and standard deviation of 80 in TIMSS. For this, one 'doubloon' country participating in both LLECE and TIMSS during roughly the same period, namely Colombia, was used. The report provides a list of 163 countries, plus five Canadian provinces, each with what is referred to as a 1965 to 2015 across-subject 'mean cognitive skills' score, mostly for the primary and secondary levels separately, and with a gender breakdown added where possible. The metric used is that of TIMSS (Altinok *et al*, 2018: 36). For the many countries participating in international assessments only during the last couple of decades, the average appearing in the list would not take into account earlier performance. There are a statistics which may appear counter-intuitive. For example, China's average score at the secondary level, 550, is virtually on a par of that of the Republic of Korea, with 553, though for some years up to 2010 the poverty

[14] Southern and Eastern Africa Consortium for Monitoring Educational Quality.
[15] Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education).

rate in China was 10%, against virtually zero for Republic of Korea[16]. The average for China is almost certainly an over-estimate, given that it is based on PISA testing occurring just in Shanghai (in 2012 and 2015), Beijing, Jiangsu and Guangdong (just 2015). These are parts of China displaying GDP per capita figures well above the national average. These types of important details regarding the tested population are not made clear in the report.

Altinok *et al* refer to a panel dataset they produced, called the 'Harmonized Learning Outcomes (HLO) database'. The HLO dataset would include percentages of students surpassing a minimum level of proficiency. This dataset was by March 2019 not available online. What is publicly available is the database produced for the UIS by Altinok (2017), where a methodology very similar to that of Altinok *et al* (2018) was followed[17].

2.5 Minimum proficiency cut-scores and calculating percentage proficient

Here the question of how minimum proficiency cut-scores are set within assessment programmes is touched on briefly. The determination of a common minimum proficiency level across programmes is discussed in an accompanying report which describes a global proficiency projection model.

How different assessment programmes set cut-scores to separate categories of performance, where each category is described in educational terms, has been summarised in a number of places: UNESCO Institute for Statistics (2017c: 15; 2017b) and Altinok (2017: 55). No programme establishes a hard dividing point between proficient and non-proficient, for the obvious reason that how one defines proficiency is to a large extent subjective and culture-specific. Thus, for instance, SACMEQ has, from weaker to stronger, categories labelled 'Emergent reading', 'Basic reading', and 'Reading for meaning' – there are further categories above this. Clearly, which of these three categories qualifies within the category 'proficient' is debatable.

A basic fact is often not mentioned, namely that some programmes have cut-scores which do not involve rounding, while some do. The two cut-scores separating the three SACMEQ categories mentioned above are 372 and 412 on the SACMEQ reading scale[18]. This suggests competency levels were defined, and then thresholds were found in the data, without any rounding. Similarly, PISA has non-rounded thresholds[19]. TIMSS and PIRLS [20] on the other hand, use cut-scores which are all multiples of 25, and exactly 75 points apart: 400, 475, 550, 625. An examination of the TIMSS technical documentation makes it clear competency descriptors were attached to an existing scale. It is not as if the existing scale was adjusted to fit competency descriptors. Whether or not cut-scores are the result of rounding clearly implies rather different approaches to the determination of cut-scores, and is something to take into account when they are interpreted.

[16] World Development Indicators of the World Bank, available at http://datatopics.worldbank.org/world-development-indicators.
[17] Database, titled 'UIS Learning Outcomes Anchored Database' available at http://tcg.uis.unesco.org/data-resources.
[18] Obtained through analysis of the 2007 SACMEQ microdata. It appears these cut-scores are not made explicit in any public document, though the SACMEQ categories have been widely used.
[19] OECD, 2016: 191.
[20] Progress in International Reading Literacy Study.

For the purposes of understanding progress, what is important is that cut-scores correspond to points in a distribution which is roughly normal. If the cut-score lies in the right-hand (better) half of the national distribution, the opportunities for improving the country's percentage proficient statistic are particularly good. Specifically, if one assumes a constant improvement in terms of the country's test score mean, the percentage of children who are proficient can be expected to improve faster if the cut-score is situated in the right-hand half of the distribution (as opposed to the left-hand half), because the curve slopes upward. These dynamics are explored in more depth in the accompanying report.

2.6 Improvements in terms of fractions of standard deviations

As demonstrated in section 2.7, it has become common to gauge improvements in education outcomes over time in terms of a fraction of a standard deviation in test scores. The method can use the standard deviation of a particular year, possibly the start or end of a series, or the mean of the standard deviations over time for a particular unit, such as a country. This method is especially attractive if one wants to compare progress measured by different assessment programmes.

Ost *et al* (2016) address an obvious question. If standard deviations are different across assessment programmes, or across countries, then are comparisons of change expressed in standard deviations really valid? Put differently, if one country is twice as unequal as another, then should one not avoid making comparisons using a metric which is a fraction of the level of inequality? Clearly, the assumption made when standard deviations are used as a measure of progress, is that standard deviations do not differ substantially.
Using PISA data, Ost *et al* (2016) conclude that differences in standard deviations across countries are large enough to warrant careful attention in across-country comparisons, when the standard deviation is the metric. Nonetheless, the utility of the standard deviation in this context is acknowledged. There seems to be no obvious alternative. Ost *et al* (2016) put forward a simple adjustment method which takes into account standard deviation differences, and which largely resolves the comparability problem. Whether standard deviations across countries are similar is explored further below (see **Figure 5**).

2.7 Existing work on business-as-usual rates and 'speed limits'

Economists have been particularly interested in how fast the quality of learning outcomes, or 'cognitive skills' as economists often put it, can improve in developing countries. This has followed findings by Hanushek and others on how very large the impact of progress in educational quality on economic growth in the long run is. It is in fact largely these findings that prompted the shift towards what children learn in the education SDGs. Projections of learning outcomes produced by economists deal with progress in average scores, not the proportion of children reaching minimum proficiency thresholds. These projections have been produced largely to quantify the long run effect of educational improvement on economic growth.

To illustrate, Hanushek and Woessman (2007: 43) attempt to answer the question: 'how fast does any [education] reform achieve its results'? Based on an assumption that seemed reasonable, namely the assumption that in a period of ten to thirty years, a middle income country could halve its quality distance from a developed country, the authors conclude that an improvement of 0.5 standard deviations over the period is possible, in the context of an 'aggressive reform plan'. This translates to 0.050 to 0.017 standard deviations a year, corresponding to the range of 10 to 30 years. A standard deviation here is the standard deviation of student scores across many countries in PISA. In PISA, but also many other testing programmes, this standard deviation is set to equal 100 points, based on actual standard deviations seen among countries participating in an initial year of the programme. The thrust of Hanushek and Woessman's argument is that educational improvement will result in an economy that is around 15% larger than in a business-as-usual scenario fifty years into the future, and around 35% as large eighty years into the future. This underscores both the critical importance of educational quality – few interventions have been found to bring about this degree of change to economic growth over the longer term – but also the patience and persistence that societies and governments need to exercise when it comes to bringing about development through educational change. Moreover, enduring educational quality reforms need to focus particularly on the start of the education process, at the primary and pre-school levels. The importance of intervening early on in the education process explains in part the unavoidable slowness of the development process. It takes over a decade for lower primary school students to enter the labour market.

Hanushek and Woessman's 2007 paper is summarised in a World Bank report from the same year[21].

In Hanushek and Woessman (2009: 21), the work is taken forward in part by examining empirically, and through imputation and the equating of different assessment programmes, what improvements have been seen in the past. The focus is on developed countries, as these countries have the longest time series of test scores. The variations across countries are large, with Netherlands, Finland and Canada improving their test score averages by around 0.01 standard deviations a year in the period 1975 to 2000, against a decline of between zero and minus 0.005 for Germany and Italy[22]. Correlations with changes in the economic growth rate are found to be high. The work in the 2009 report is reproduced in an OECD (2010) report.

Outside of the Hanushek-Woessman orbit, Mourshed *et al* (2010: 16), in an influential McKinsey report, use trends from twelve countries or regions they consider noteworthy improvers, to arrive at an improvement of 0.115 standard deviations in ten years, or 0.012 per year. The twelve can be considered advanced economies.

In a 2014 doctoral dissertation by the author of the current report, Gustafsson (2014: 134), arrives at the conclusion that an optimistic though feasible policy target could be premised on a test score improvement of 0.06 standard deviations a year. This figure is derived from trends seen in several testing programmes in the years 2000 to 2009, and specifically the trends of eleven relatively fast improvers, of which six are middle income countries, and the remaining five high-income OECD countries. It is also concluded that developing

[21] Hanushek and Wößmann, 2007.
[22] These values are off the horizontal axis of Figure 2 in the report. They are also implied by Figure B3.

countries tend to have more scope for rapid improvements, given their distance from what one can think of as natural ceilings for cognitive skills.

Altinok *et al* (2018: 33), using their harmonised scores, confirm the trends for countries such as the Netherlands and Germany referred to above[23]. For countries outside the set of high-income OECD countries, annual improvements over 35 years as high as 0.03 standard deviations a year, for Hong Kong, and 0.02, for Iran, are found. Iran's trajectory is interesting, given that the country is a large developing country, and that the time period of 35 years is exceptionally long. However, the improvement figure of 0.02 cannot be replicated using what would represent at least a part of the data used by Altinok *et al* (2018). Official TIMSS reports include graphs with country trends over time. The official TIMSS Grade 9 trend for Iran extends from 1995 to 2015, with six points in time, an exceptionally good availability of data for a developing country[24]. This official trend produces an annual slope of 0.44 TIMSS points, which translates to around 0.0044 standard deviations a year, not the 0.02 reported in Altinok *et al* (2018: 33). The dataset published with Altinok (2017) suggest that TIMSS Grade 9 values for the period 1995 to 2015 were indeed the source used in the case of Iran.

There seems to be little work on this topic beyond what is discussed above. This can be considered a serious gap from an education planning perspective, especially now that the focus globally has moved squarely towards improvements in learning outcomes. There are several possible explanations as to why more work to guide the setting of targets has not emerged in recent years. The preference in many national education authorities may be for targets in an area as sensitive as learning outcomes to be set politically, and not empirically. The very ambitious SDGs can be said to reinforce such an approach. In economics of education, much of the focus on improvement has been on what is possible within controlled 'experiments' consisting of a random sample of schools. This an area which lends itself to empirical research. The matter of how whole education systems improve, a more complex area requiring a more inter-disciplinary approach, has been left largely untouched.

It should be emphasised that results from experimental research are not simply transferable to whole schooling systems. This is often acknowledged by the producers of this research. The results from experimental research are often encouragingly good, quite often as high as 0.2 of a standard deviation following an intervention of one or two years[25]. This represents a larger gain than the 0.06 of a standard deviation referred to above, found by gauging the trends of whole countries. Results from the experimental research are difficult, some would say impossible, to replicate successfully across an entire system for several reasons, including 'general equilibrium effects'. To illustrate the latter, while teacher unions may not oppose a small research intervention, they might want to alter or even stop the same intervention when taken to scale.

[23] For instance, a 50-score gain over 50 years for the Netherlands, comes about 0.01 standard deviation a year, given the TIMSS scale used by the authors.
[24] Mullis *et al*, 2016: Exhibit 1.6.
[25] McEwan, 2015.

To conclude this section, the little work that has been undertaken to establish how fast schooling systems improve points to a 'speed limit' of anywhere between 0.01 of a standard deviation a year – for a developed country over 25 years – to 0.06 standard deviations a year – for countries of a variety of development types, but for a shorter period of a decade. Below these 'speed limits', it is quite possible for countries, such as Germany and Italy between 1975 to 2000, to experience what can be termed a business-as-usual trend of zero or even slightly negative growth in learning outcomes.

3. New analysis of change in terms of means and percentage proficient

This section presents new analysis aimed at bringing more certainty to questions around historical rates of educational improvement. It therefore builds on past work described in section 2.7. The focus is on mean scores, but also on what has received virtually no attention: trends with respect to the percentage of students who are proficient. Improvements in the mean are expressed relative to standard deviations, given that the three testing programmes considered, PIRLS, PISA and LLECE, each has its own score metric. However, as will be explained, this use of the standard deviation introduces its own bias.

3.1 Evidence from PIRLS

The analysis begins by examining trends in the Progress in International Reading Literacy Study (PIRLS) programme. PIRLS is particularly important insofar as it focusses on reading at the lower primary level, specifically Grade 4. Of all the SDG 4.1.1 sub-indicators, the one focussing on lower primary reading is perhaps the most important, given that reading is a foundation for everything else in schooling.
For the graphs that follow, PIRLS trends for 28 countries were used. Of these, 22 are outside of the group of high-income OECD countries. The definition of 'high-income' was conservative, however, as it required the country to have been classified as high-income by the World Bank already in 2012[26]. Given that the intention was largely to understand trends in developing countries, the only PIRLS participants in the high-income OECD group used in the analysis were the six who are also G7 members: United States, Canada, France, Germany, Italy, United Kingdom (Japan has not participated in PIRLS). England's results were used to represent the United Kingdom.

Table 2 underlines what a relatively small 'sample' the 28 selected PIRLS countries are. They represent 15% of the world's children.

[26] The 2012 *World Development Report* (World Bank, 2011) used. Since then, some countries, such as Latvia and Chile, have emerged as high-income countries in some years.

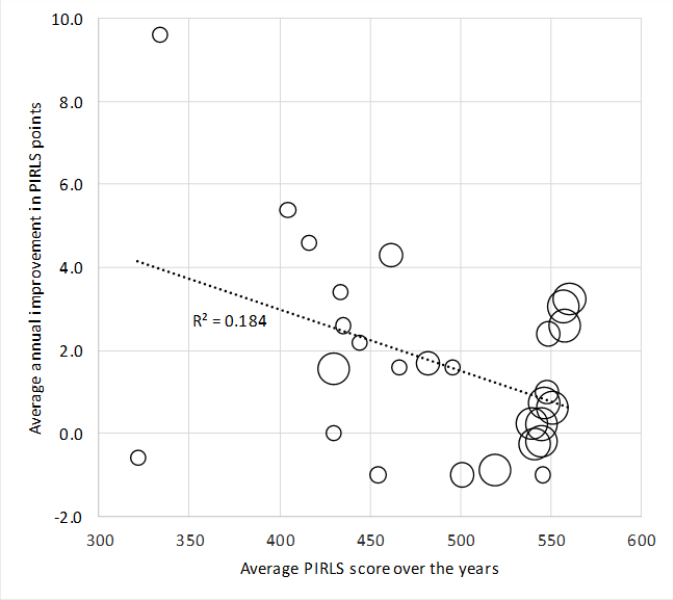
Table 2: Extent of PIRLS

UNESCO 2019 region	PIRLS countries in the analysis	% of all countries	% of child population
Sub-Saharan Africa	1	2	4
Northern Africa and Western Asia	7	30	15
Central and Southern Asia	1	7	3
Eastern and South-eastern Asia	4	21	17
Oceania	0	0	0
Latin America and the Caribbean	2	4	7
Europe and Northern America	13	26	73
Total	28	12	15

F**igure 1** below illustrates three variables. The size of each marker reflects the number of years for which an average score for the country existed, the range being from two to four years, and the four years being 2016, 2011, 2006 and 2001. The horizontal axis represents the average PIRLS score across all years. The vertical axis represents the annual change in the average score, in the sense of the annual slope across the two, three or four data points. Underlying statistics are as they appear in the official PIRLS reports for the four years[s27].

[27] Mullis *et al*, 2017; Mullis *et al*, 2012; Mullis *et al*, 2007; Mullis *et al*, 2003.

Figure 1: Score increases in PIRLS 2001-2016



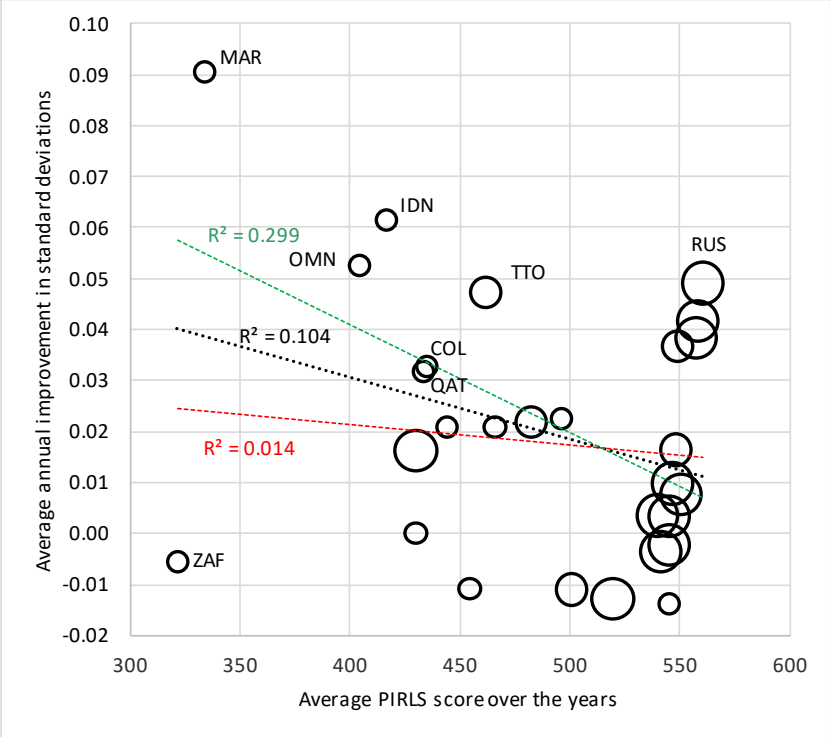
Note: The three sizes of the markers represent (from smallest) two, three and four years of data in the series. The trendlines here and in the following graphs are not weighted by the size of each marker.

Figure 2 below is similar, but gets closer to what is needed. Here the vertical axis represents the annual change in terms of standard deviations, where the standard deviation used for each country is the average across the two most recent ones.

The two graphs confirm that most countries in PIRLS are characterised by an improvement: of the 28 countries, 20 experienced a positive slope. Moreover, the black regression line suggests that countries starting off from a lower base are likely to experience larger gains. However, the evidence for this is not very strong. In the official PIRLS reports, Morocco (MAR) is considered to be such a weak performer that its scores in both years, 2011 and 2016, are said to suffer from limited reliability. The PIRLS tests were not designed for a country with such weak performance. Removing Morocco from the calculation of the regression line results in the alternative red trendline, which is virtually flat, suggesting that less developed countries do not experience larger improvements. How South Africa (ZAF) is understood makes a difference. The author of the current report has argued that if one considers South Africa's older PIRLS 2006 results (something the official PIRLS reports do not do), South Africa displays an annual improvement of 0.06, and not roughly zero as suggested by Figure 2[28]. This, if we also include Morocco, would result in the green trendline seen in the graph, which points fairly strongly to greater improvements for countries with a worse baseline.

[28] Report titled *TIMSS, SACMEQ and PIRLS: Data issues*, available online on the author's blog.

Figure 2: PIRLS improvements in terms of standard deviations

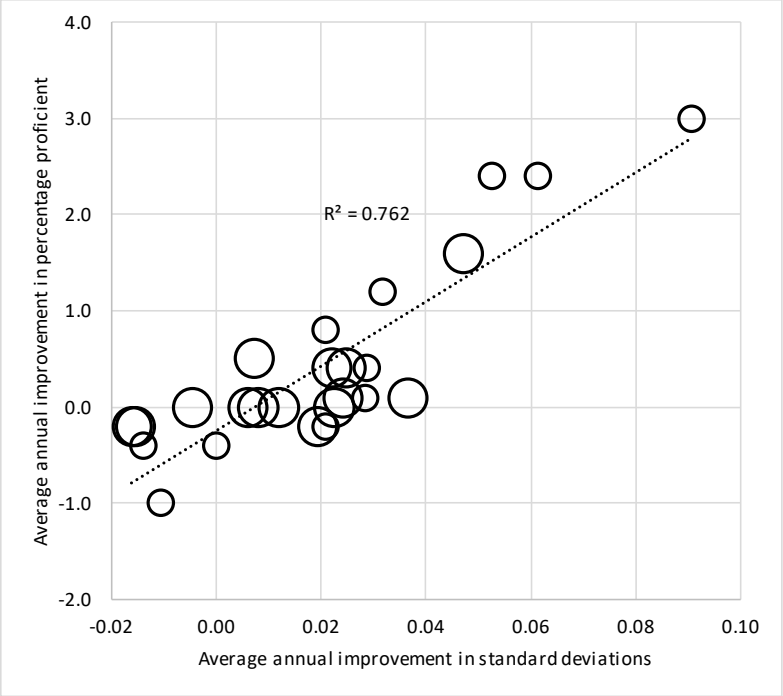


One could conclude from the above that an optimistic business-as-usual annual improvement would follow roughly the black trendline. This would be optimistic in the sense that we can assume that PIRLS participants are countries with relatively well-organised schooling systems. This is why they prioritise participating in PIRLS. A system which does not pay much attention to educational improvement is not likely to participate in PIRLS. Importantly, there are several PIRLS developing countries which have experienced improvements beyond 0.02 standard deviations a year: Oman (OMN), Indonesia (IDN), Trinidad and Tobago (TTO), Colombia (COL) and Qatar (QAT). It seems safe, even conservative, to say that a global 'speed limit' lies approximately where the green trendline appears. There are enough countries, from a variety of world regions, displaying such improvements for it to be defensible to say others can be expected to emulate this.

The relatively fast improvement of a few high-performing countries is noteworthy: Russia (RUS), but also Hong Kong, Singapore and Chinese Taipei. Yet the average annual improvement for all countries in Figure 2 scoring above 500 on the horizontal axis is just 0.012, against 0.029 for countries scoring below 500. Expecting more from educationally less developed countries in the form of higher rates of improvement, or higher 'speed limits', seems justified.

Figure 3 confirms that what should be happening is happening: improvements in the mean, expressed as a proportion of the standard deviation, are strongly correlated with improvements in the percentage of students reaching a minimum proficiency level, in this case the PIRLS 'low international benchmark' of 400.

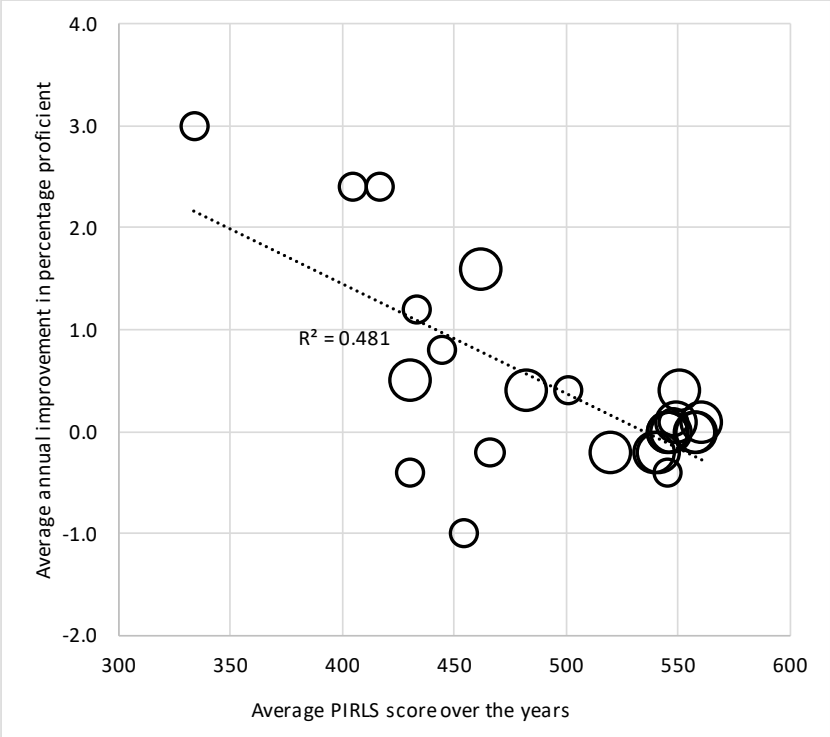
Figure 3: PIRLS improvements in means and percentage proficient



Note: This graph uses 25 countries as comparable percentage proficient statistics are only available in the PIRLS reports of 2006, 2011 and 2016. The 2001 report uses a different minimum threshold.

Figure 4 below, like Figure 2, focusses on the rate of improvement relative to the general level of performance, except here the vertical refers to percentage proficient. The downwardly sloping trendline, with a relatively high R2, obtained using data from the 25 countries for which the analysis was possible, confirms less developed countries can expect greater improvements. The graph suggests that an annual improvement of up to 2.0 percentage points a year for the weakest performers has been happening in recent years, at least among PIRLS participants. The fact that all developed countries should display an annual improvement of around zero should not come as a surprise. These countries are left with very few non-proficient children. The 2016 average across the six G7 countries included in the graph is 4% non-proficient, the range being 2% to 5%. It is likely to be particularly difficult to eliminate this 'last mile' standing in the way of achievement of a percentage proficient value of 100%.

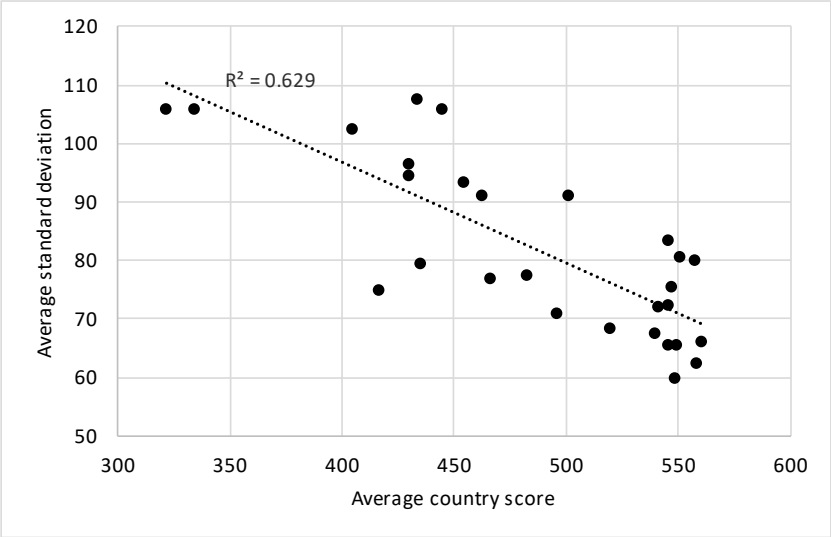
Figure 4: PIRLS percentage proficient business-as-usual trends



The average rate of improvement across all 25 countries displayed in the above graph is 0.4 percentage points a year.

In one respect, the above analysis under-estimates the gains seen in developing countries. This is because one important fact is ignored: the standard deviations used to gauge improvements are larger among worse performing countries. This is seen in **Figure 5** below. This is clearly worth taking into consideration in any modelling of future trends.

Figure 5: Standard deviations in PIRLS



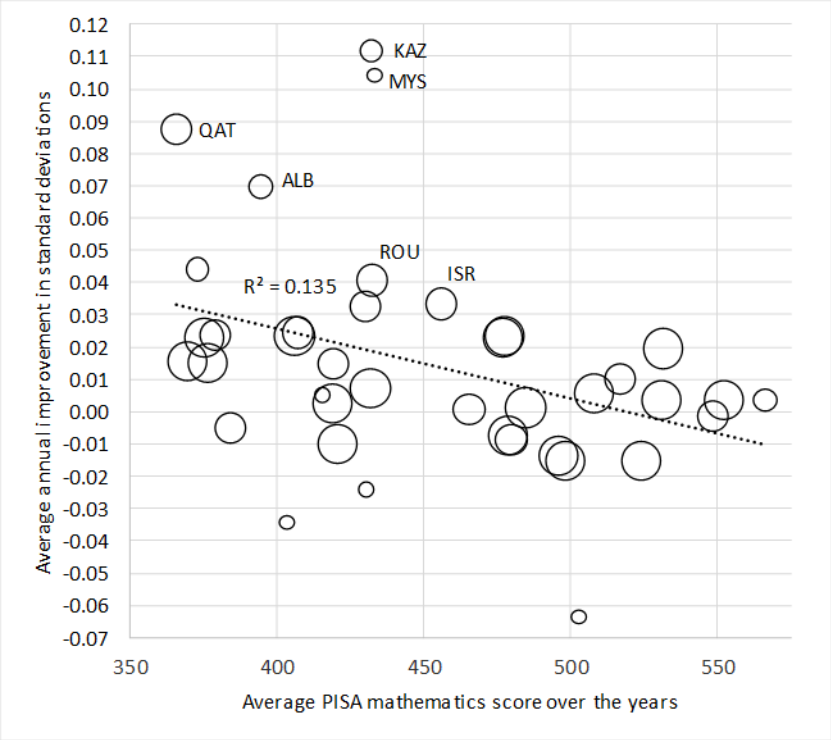
Note: Standard deviations are the mean across the two most recent years available, the source being the official PIRLS report. The means are calculated over up to four years, as for the earlier graphs.

3.2 Evidence from PISA

Figure 6 below is like the earlier Figure 2, except here PISA trends are analysed. Again, only G7 countries from the set of high-income OECD countries were considered. The online PISA Data Explorer[29] was used to obtain mathematics test score means for 39 countries for up to five years: 2003, 2006, 2009, 2012 and 2015. Standard deviations used to calculate the vertical axis are those of 2015[30].

[29] https://pisadataexplorer.oecd.org/ide/idepisa.
[30] From OECD (2016: 389).

Figure 6: PISA improvements in terms of standard deviations



Note: The four sizes of the markers represent (from smallest) two, three, four and five years of data in the series.

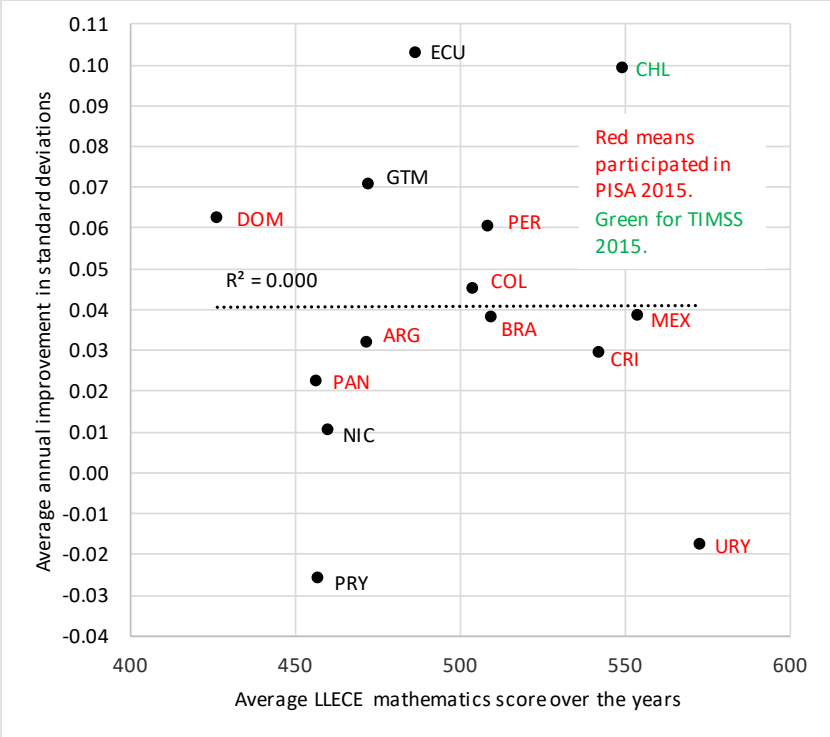
The picture is not that different to the one obtained using PIRLS data. Most countries, 28 of the 39, experienced positive growth with respect to the mean score. This growth was higher among countries with lower test means, in other words the least developed countries. In fact, this is more pronounced in the PISA graph, which lacks fast improving countries on the right-hand side of the graph. The suggested magnitude of BAU and 'speed limit' improvements are not very different in the PISA and PIRLS graphs.

3.3 Evidence from LLECE

For the Latin American LLECE assessment programme, the only comparisons over time available are those between the 2006 and 2013 waves of LLECE, known respectively as SERCE and TERCE. The improvement patterns for the 14 countries with values in both years in Grade 6 mathematics are shown in **Figure 7** below. All but two countries experienced improvements. However, the pattern seen earlier of greater improvements among lower performing countries is not seen here. This would in part be because LLECE covers just developing countries. The average improvement is 0.04 standard deviations, not far from the 0.033 annual improvement seen over ten years for Colombia in PIRLS – see Figure 2 above[31].

[31] Colombia's two points in that graph are from 2001 and 2011.

Figure 7: LLECE 2006-2013 improvements in terms of standard deviations



Note: Points represent progress in Grade 6 mathematics. Improvements in terms of LLECE scores are from UNESCO (2014: 41). Standard deviations were derived from the TERCE microdata available at http://www.unesco.org/new/en/santiago/education/education-assessment-llece. Standard deviations corresponding to a scale where 500 (not 700) is the mean were calculated.

The LLECE data do not point to any clear correlation between average score, or degree of improvement, on the one hand, and participation in one of the large programmes TIMSS or PISA, on the other. However, this may not say much about the suspicion raised earlier, namely that countries participating in no international assessment at all could be exceptionally weak performers, even relative to other countries with similar GDP per capita values.

4. Concluding remarks

Though SDG indicator 4.1.1 focusses on improvements in terms of the percentage of proficient children, most of the limited literature on how fast countries improve has focussed on improvements in country mean scores. As is shown in the report and projection tool accompanying the current report, basing indicator 4.1.1 projections on trends in the mean score, and translating these to trends in the percentage proficient, appears to be a logical approach. Percentage proficient statistics can be difficult to engage with directly, because their future trajectory depends on the shape of the distribution of children considered non-proficient, who must become proficient in future.

The existing analyses discussed in this report point to annual improvements in the mean of between 0.01 and 0.06 standard deviations a year. As discussed above, while fractions of a standard deviation should be used with caution, they are nevertheless a useful device to broaden the analysis to several assessment programmes at a time. The examination of PIRLS, PISA and LLECE trends presented in the report points to trends which are roughly within this 0.01 to 0.06 range. Specifically, the data support a business-as-usual (BAU) trend ranging from 0.04 standard deviations gain a year for the least educationally developed countries, and zero to 0.01 for developed countries. The fastest improvements witnessed in the three programmes in the case of developing countries are around 0.08 standard deviations a year.

Two types of analysis would further strengthen the knowledge base which can be used to inform targets and historical analyses into whether countries were succeeding or not. One is more frequent and rigorous interrogation of the trends of individual countries. For example, do the underlying data support the exceptional improvements experienced by Qatar in PISA, or Indonesia in PIRLS? In particular, is there nothing to suggest that the sample may have shifted? Despite the application of rigorous standards in the programmes in question, sampling problems have been known to distort trends. Confirming that published trends are real is important not just for the countries concerned, but for all countries, who need to know what is possible.

Secondly, while research into the impacts of specific interventions in samples of schools have generated valuable knowledge about educational change, what is still too under-explored, is how whole countries improve. For instance, assuming Trinidad and Tobago has really improved by almost 0.05 standard deviations a year, across three waves of PIRLS, what are the education policy, political, social and economic factors behind such a trend? Answering such questions is necessarily a multi-disciplinary task, yet an important one.

References

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

Altinok, N, Angrist, N & Patrinos, H.A. (2018). *Global dataset on education quality* (1965-2015). Washington: World Bank.

Carnoy, M., Khavenson, T., Fonseca, I. & Costa, L. (2015). *Is Brazilian education improving? Evidence from Pisa and Saeb*. *Cadernos de Pesquisa*, 45(157).

Ethiopia: Federal Ministry of Education (2015). Education Sector Development Programme V. Addis Ababa.
Gustafsson, M. (2014). *Education and country growth models*. Stellenbosch: University of Stellenbosch.

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

Hanushek, E.A. & Woessmann, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation*. Washington: National Bureau of Economic Research.

Hanushek, E.A. & Wößmann, L. (2007). *Education quality and economic growth*. Washington: World Bank.

Hawaii: Department of Education (2003). *Reading and mathematics AYP starting points, intermediate goals, annual measurable objectives*. Honolulu.

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

Jerrim, J. (2013b). *People having a pop at PISA should give it a break*. London: University of London. Available from: <http://ioelondonblog.wordpress.com/2013/07/30/people-having-a-pop-at-pisa-should-give-it-a-break/> [Accessed August 2014].

Klein, R. (2011). *Uma reanálise dos resultados do Pisa: problemas de comparabilidade*. Rio de Janeiro: Avaliação e Políticas Públicas em Educação.

McEwan, P.J. (2015). *Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments*. Review of Educational Research, 85(3): 353-394.

Mourshed, M., Chijioke, C. & Barber, M. (2010). *How the world's most improved school systems keep getting better*. New York: McKinsey & Company.

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: Boston College.

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

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. & Kennedy, A.M. (2003). *PIRLS 2001 international report*. Chestnut Hill: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P. (2007). *PIRLS 2006 international report*. Chestnut Hill: Boston College.

OECD (2010). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. Paris.

OECD (2016). *PISA 2015 results: Excellence and equity in education: Volume 1*. Paris.

Ost, B., Gangopadhyaya, A. & Schiman, J.C. (2016). *Comparing standard deviation effects across contexts. Education Economics.*

Sandefur, J. (2016). *Internationally comparable mathematics scores for fourteen African countries.* Washington: Center for Global Development.

UNESCO Institute for Statistics (2017a). *Metadata for the thematic and global indicators for the follow-up and review of SDG 4 and Education 2030.* Montreal.

UNESCO Institute for Statistics (2017b). *SDG reporting: Linking to the UIS reporting scale through social moderation*. Montreal.

UNESCO Institute for Statistics (2017c). *Exploring commonalities and differences in regional and international assessments.* Montreal.

UNESCO (2014a). *Reporte técnico: Tercer Estudio Regional Comparativo y Explicativo*. Santiago.

UNESCO (2014b). *Education for All global monitoring report 2013/4: Teaching and learning: Achieving quality education for all.* Paris.

UNESCO (2016). *Global Education Monitoring Report 2016: Education for people and planet.* Paris.

UNESCO (2017). *Global Education Monitoring Report 2017/18: Accountability in education: Meeting our commitments.* Paris.

United Nations (2015). *Resolution adopted by the General Assembly on 25 September 2015: Transforming our world: the 2030 Agenda for Sustainable Development*. New York.

United Nations (2017). *Revised list of global Sustainable Development Goal indicators*. New York.

World Bank (2011). *World Development Report 2012: Gender equality and development*. Washington.

World Bank (2018). *World Development Report 2018: Learning to realize education's promise*. Washington.