RISE

RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

WORKING PAPER
January 2018

Worldwide Inequality and Poverty in Cognitive Results: Cross-sectional Evidence and Time-based Trends

Luis Crouch and Martin Gustafsson

RISE-WP-18/019

Funded by:
UKaid from the British people
Australian Aid

Oxford Policy Management

Blavatnik School of Government, University of Oxford

Center for Global Development

Worldwide Inequality and Poverty in Cognitive Results: Cross-sectional Evidence and Time-based Trends[1]

Draft: 8 January 2018

Luis Crouch[2]
Martin Gustafsson

[1] This is a companion note to Crouch and Rolleston (2017), a RISE Programme Insight note. It delves into the issues in more detail than was possible in that note.

[2] RTI International and University of Stellenbosch, respectively. Crouch is a member of the RISE Programme's Intellectual Leadership Team.

Contents

1. Introduction

The Sustainable Development Goals (SDGs) for education represent a major departure from the Millennium Development Goals (MDGs) - at least if educational leaders act seriously in their pursuit - in at least two important respects.

First, the goals now pertain to learning outcomes.[3] Out of eleven indicators, four focus directly on learning or developmental outcomes, and, in addition, one of the eleven goals (4.7) would require measurement of learning outcomes, even though the relevant indicator currently does not explicitly mention learning outcomes. In presentational terms, the lead indicator (the one presented first in the list) is all about learning and is really composed of three sub-indicators. This is a major departure from the previous set of global goals, the MDGs.[4] These had only one indicator which could arguably be said to pertain to learning outcomes produced by an education system in relative "real time," namely the literacy rate of 15- to 24-year-olds. However, measurement was often based on self-reports in surveys or censuses, or was based on imputations derived from completed years of education. The headline indicators for the MDGs were really all about access. The formal indicator expressed in the MDG list was completion of primary schooling, but the discussion among global educational leadership focused almost entirely on the number of children out of school and on gender differences in these numbers.[5]

The debate started to shift in the early and mid-2000s. The Global Monitoring Report for the Education for All movement, produced in 2012, estimated that some 250 million children were in school but essentially learning nothing. This got a large amount of coverage. The implicit comparison against some 60 million children out of school was sobering. In addition, scholars, non-governmental organizations (NGOs), and development agencies started calling for something akin to "Millennium Learning Goals." Perhaps more telling than the absolute-amount estimate of 250 million hardly-learning was Filmer, Hasan, and Pritchett's (2006) early estimate that the median student in poor countries learns at about the same

[3] See list at https://sustainabledevelopment.un.org/sdg4, see tab on "Targets & Indicators" [Accessed on 25 August 2017].
[4] See list at http://www.unmillenniumproject.org/goals/gti.htm#goal2 [Accessed on 25 August 2017].
[5] Number of Google hits hardly represents the joint view of global education policy makers, though it might represent the "buzz" that feeds them or that they feed. In any case, a search for ["out of school children" "millennium development goals"] produces 217,000 hits. One for ["primary school completion" "millennium development goals"] produces only 17,500 hits. Global commissions and personages typically focused on the absolute number of children out of school. A search for ["Ban Ki Moon" education "out of school children" "millennium development goals"] produced 117,000 hits, whereas one for ["Ban Ki Moon" education "out of school children" "millennium development goals"] produced 2,700 hits. Officials of governments and development agencies did focus more on the completion rate.

level as children around the 5th percentile in OECD countries. Thinking in terms of percentiles suggests that the median (or below) child in poor countries would be a candidate, or a near candidate, for disability-related support in the OECD—a sobering perspective.

The focus on learning is more demanding, but also more meaningful, than a focus on mere access to schooling, because there is far more inequality in learning around the world, than there is in access to schooling. As an example, consider reading skills. A crude but serviceable index of worldwide inequality of this indicator, just within the PISA 2015 reading database, can be crafted by taking the percentage of students performing at or above an acceptable minimum level (Level 2 in PISA 2015 reading), taking the range of this percentage between the three highest and three lowest-performing countries, and dividing this by the median performance. That produces an index value of 0.81. Now, taking two indicators of access, namely the gross enrolment ratios for secondary and tertiary education (the next two access frontiers), a similar calculation yields 0.30 and 0.57 respectively, for the same countries. In a loose sense, and using a database that does not include the worst-off countries in the world, the inequality in learning outcomes is 170 percent greater than the inequality in access to secondary education, and 43 percent greater than the inequality in access to tertiary education.[6]

Second, there is a great deal of focus on inequality in the SDGs. In the MDGs, since the goal was in any case 100 percent access, and access is a binary issue (one is either in school or not, one completes primary school or not[7]), a goal of equality is implicit in a goal of 100 percent access. In addition, much of the declarative emphasis was on gender inequality. However, gender inequality is a relatively weak proxy for overall inequality. The SDGs are different. They emphasize inequality as created by quite a few factors (gender, region, income, disability, etc.). (Though they do not cover what one might call "pure" or "total" inequality, that is, the total dispersion in scores, due to factors such as income and region, but also, importantly, due to lack of quality assurance and standards.) This is an important lack in the SDGs, but researchers and policy-makers are likely to pay attention to the issue anyway. Moreover, learning outcomes are a much less binary and more complicated

[6] Just to take a data point from *within* a developing region, in Latin America's TERCE Grade 6 reading results, the cognitive gap between the top two and bottom two performers, measured via the same index as in PISA 2015 reading (but in TERCE taking the top two performance levels, as the TERCE levels are much "easier" than PISA's), is 1.25, whereas the gaps in access to secondary and tertiary education respectively, are 0.31 and 0.53, respectively: the cognitive gap is 4 times larger than the gap in access to secondary education, and 2.3 times larger than the gap in access to tertiary education.

[7] Admittedly, this is a bit of an over-simplification. Poor countries are plagued by both pupil and teacher attendance issues, for instance.

phenomenon than access to schooling, and therefore so is the measurement of its inequality.

Taking note of this new dual emphasis of the SDGs, this paper assembles the largest database of learning outcomes inequality data that we know of, and explores key issues related to the measurement of inequality in learning outcomes, with a view to helping countries and international agencies come to grips with the key dimensions and features of this inequality. Two issues in particular are explored. First, whether, as countries improve their average cognitive performance (as measured by international learning assessments) from the lowest to middling levels, they typically reduce cognitive skill inequality or, more importantly perhaps, whether they reduce absolute lack of skills. Second, whether most of cognitive skills inequality is between or within countries. In dealing with these measurement issues, the paper also explores the degree to which measures of cognitive skills are "proper" cardinal variables lending themselves to generalizations from the field of income and wealth distribution—the field for which many measures of inequality and its decomposition were first applied. To do this, we look into whether using the item response theory (IRT) test scores of programmes such as TIMSS influence these types of findings, relative to the use of the underlying and more intuitive classical test scores. Patterns emerging from the classical scores are far less conclusive than those of the IRT scores, in part due to the greater ability of the IRT scores to discriminate between pupils at the bottom end of the performance spectrum. An important contribution of the paper is to examine the sensitivity of standard measures of inequality to different sets of test scores. The sensitivity is high, and the conclusion is that meaningful comparisons between test score inequality and, for instance, income inequality are not possible, at least not using the currently available toolbox of inequality statistics. Finally, the paper explores the practical use of school-level statistics from the test data to inform strategies for reducing inequalities.

2. Brief literature review

The exploration of inequality in educational results, in a manner that borrows from the analysis of economic inequality, is not new. Economists and educationalists have, for some time, explored the generalization of concepts such as the Gini coefficient, and other measures of inequality, whether global (between-country inequality) or local (within-country inequality), to education. However, typically these generalizations have been used for measures of access or of spending, not for measures of learning outcomes. In economics, a key paper outlining findings germane to those here is Dollar and Kraay's (2001) "Growth is good for the poor." The paper makes the argument that since the share of income of the

poorest quintile of the populations they study (over many decades) does not decrease when average incomes increase, incomes for the poor must be increasing. The authors find some evidence that one reason that growth is good for the poor is that otherwise pro-growth policies, such as inducing greater educational attainment (but not necessarily *spending* more on education), benefit the poor disproportionately. But this evidence pre-dates the availability of the sorts of data we present here, so the conclusions are somewhat equivocal. Our paper makes a similar case, though the logic is somewhat flipped on its head: we argue that average levels of cognitive achievement in countries seem to increase most, from the lowest average levels to middling average levels, when the average levels of cognitive achievement of those who, at baseline, had the lowest absolute levels of achievement, grow the fastest. One point where the analogy between income and knowledge breaks down is that knowledge cannot be redistributed ex-post.[8] In some sense, though, this makes the *creation* of human capital an even more attractive (and urgent) way to generate equality. Sahn and Younger (2007) also point this out. Our paper would also appear to have something to say about the discussion (e.g., Milanovic 2012) on whether "class" or "location" has become more or less important over time in driving income distribution or, to put it in other terms, whether inequality is growing within or between countries. We will conclude that today, "about half" of inequality is internal to countries. But we cannot conclude anything about how this has changed over time, as data on learning achievement are not available for decades and decades, as data on income and wealth distribution are. All we can say, as will be seen below, and with many caveats, is that somewhat more than half of cognitive inequality is within countries.

Note that this paper does not delve into the issue of income inequality per se, nor on the relationship between income inequality and educational inequality. Those would be long and complicated debates. Our aim is more modest, and it is simply to establish some facts about educational (and cognitive levels at that) inequality.

In education, one of the earliest examples of this kind of work, which nonetheless refers to three earlier cases, is Thomas, Wang, and Fan (2001) (associated with the World Bank) who apply the concept to years of educational attainment. Though they do not present it, one can infer that the median value for this education Gini was about 0.4, which is in the same neighbourhood as the (current) median global income Gini of 0.38. They establish certain interesting facts, such as that the Gini coefficient for years of attainment shows quickly reducing inequality the higher the average years of schooling attainment. This is sensible,

[8] Thanks to Justin Sandefur for pointing this out.

since individuals are unlikely to pile PhD upon PhD, the way they typically ambition income. "Years of education" has a reasonable upper limit, so the higher the average, the lower the inequality. It seems unlikely that one would observe such neat relationship between the inequality of learning and average levels of learning, at least using current assessments, since, as discussed elsewhere in this paper, current assessments do not seem to top out much.

The World Bank is also associated with having popularized the methodology of benefit-incidence analysis, which estimates the relationship between resources allocated to levels of education (primary, secondary, etc.) and the income groups attending those levels, hence allowing an estimation of education resources going to those groups. These analyses typically combined the use of administrative data with household survey data. Bank researchers also popularized the analysis of levels-of-education-completed by income group, using household surveys. These methodologies drew attention to the fact that the allocation of resources to tertiary education in most developing countries typically favors the already-wealthy. Benefit-incidence analyses and the development of profiles of completion by income group gained widespread attention and became almost a cottage industry for Bank education sector analyses, based on Bank-produced manuals (see, for instance, Demery 2000). Non-Bank researchers applied the methodology as well, as in Crouch, Lavado, and Gustafsson's (2009) estimates of inequality of resource distribution, and its dynamics, in Peru and South Africa.

Recently there has been an increasing interest from researchers, perhaps spurred by the growing attention to learning and the quality of education, and the availability of data, in trying to measure inequality in learning outcomes distribution. Issues such as whether inequality, using various indices, is growing or not (e.g., whether inequality increases or decreases, in a cross-section sense or over time) as mean country performance increases (e.g., Freeman, Machin, and Viarengo 2011; Oppedisano and Turani 2015; Micklewright and Schnepf 2006; and Bruckauf and Chzhen 2016), how global inequality is distributed between and within countries (e.g., Sahn and Younger 2007), and what are its determinants (e.g, Ferreira and Giroux 2011), have all seen a somewhat growing literature. The RISE Programme, publisher of this Working Paper, has a growing interest in these issues: this paper is a companion piece to the recently published note by Crouch and Rolleston (2017) which looks at many of these issues using the same range of international learning assessments as this Working Paper, but in less detail.

3. Do countries improve average learning levels by reducing inequality or absolute cognitive lacks, or both?

In this section we explore whether countries seem to improve the average cognitive level of their students while typically also reducing inequality and (the learning outcomes equivalent of inequality or relative poverty) and/or by reducing absolute cognitive lack (the learning outcomes equivalent of reducing the percentage of a population under a poverty line).

We take two approaches. First, we use published (in some cases electronically published) and "macro" (country-level) data from PISA 2015, TIMSS 2015, PIRLS 2011, TERCE, and SACMEQ III (2007) to provide an approximate but, we feel, fairly convincing answer to this question. One could wonder whether the conclusions derived from this analysis might be somewhat affected by whether one can consider learning outcomes data, as produced by these assessments, to be cardinal. Or, it may also be that micro data on the distribution of individual outcomes within countries could change our perspective. Thus, in a second step, we analyse the micro, student-level data from two of these assessments, namely TIMSS 2003, 2007, 2011, and 2015 Grade 4 mathematics, and SACMEQ III mathematics (Grade 6) to see whether the conclusions seem to hold. Using the micro data also allows us to compare results derived from Item Response Theory (IRT) to results derived from Classical Testing Theory (CTT), and to address some issues pertaining to the cardinality, or otherwise, of IRT-based scores.

3.1 Evidence from macro data

Along with Freeman, Machin, and Viarengo (2011), Micklewright and Schnepf (2006) and Bruckauf and Chzhen (2016), we calculate cross-country correlations between mean results and a measure of inequality of results as a way to judge whether higher mean performance is associated with lower or higher inequality of performance. Whereas these (and others) have calculated these measures using one or two particular assessments (or the same assessment in two years), we present here (Table 1) the correlations for the most recent round of *all* the better-known international assessments (at all grades and for all subjects).

Table 1. Correlations between mean performance and standard deviation of performance across countries			
PISA 2015 Reading	0.14	PIRLS 2011	-0.74***
PISA 2015 Mathematics	0.38***	SACMEQ III Reading	0.64*
PISA 2015 Science	0.57***	SACMEQ III Mathematics	0.76*
TIMSS 2015 Grade 4 Mathematics	-0.69***	TERCE Grade 3 Reading	0.28
TIMSS 2015 Grade 8 Mathematics	-0.29	TERCE Grade 3 Mathematics	0.39
TIMSS 2015 Grade 4 Science	-0.78***	TERCE Grade 6 Reading	0.55*
TIMSS 2015 Grade 8 Science	-0.67***	TERCE Grade 6 Mathematics	0.88***

Note: for TERCE we could not find the standard deviations for individual countries, so we used the country-specific differences between the scores at the 34th and 66th percentiles as a rough proxy for the standard deviations.[9] Furthermore, that difference had to be estimated from a table showing the scores at the 10th, 25th, 75th and 90th percentiles. The fit between scores at those percentiles, and the percentiles, was about 0.99 for all countries.

* p value less than or equal to 0.05

** p value less than or equal to 0.01

*** p value less than or equal to 0.001

The results are at best ambiguous. Some of the correlations are positive (higher mean results go along with higher inequality), some are negative. Many are statistically significant. Somewhat worryingly, the results vary greatly by assessment, and the negative correlations come from one single organization, which could lead one to think that there is something about the way the assessments are constructed or scored that generates the results. It may be interesting to compare these results to results in the literature. Freeman, Machin, and Viarengo (2011) find that the higher the median, the lower the inequality, using PISA 2009 and PISA 2000. This is *not* confirmed here, for PISA 2015 (in all three subjects). Micklewright and Schnepf (2006) use an average of results in countries that participated in

[9] These were the percentiles which, applied to the distribution of scores for all countries as a whole, best reproduced the standard deviation for all countries as a whole (100).

all assessments, and an average of the inequality of results. They find a -0.55 correlation between inequality of results and average results, so they conclude that, “There is a reasonably clear pattern: in broad terms, within country differences are highest (positive z-scores) where *average* [across assessments] achievement is lowest (negative z-scores).” (p. 10, our emphasis, our insert). But because they average many of these things out, it is difficult to tell in which assessments there is a negative versus positive correlation between inequality of results and average results. They present a table that shows that the average inter-assessment correlation in the measure for countries’ inequality results is only 0.55. And, there are only seven out of twenty-one countries where all assessments place the country as either above or below the median in inequality. So, clearly, some of the assessments measure some countries as relatively unequal and other assessments measure them as relatively equal. Thus, it is understandable that, as in Table 1, the results might be inconsistent. We have presented the results individually, assessment by assessment, and for every one of the most widely-used assessments, so one can judge clearly the ambiguity of the results.

But this does not say anything about whether increases in averages are more correlated with a reduction in the proportion of students who fall below a minimum, or an increase in the proportion of students rising above a relatively high benchmark. We propose and demonstrate an approach that allows us to address this issue using a notion akin to income poverty (as expressed in terms of percentage of population below a poverty line) as opposed to inequality.

There does not seem to be much literature on this issue. Our search turned up only one paper (Bruckauf and Chzhen 2016) that explores the issue in any depth, but it does so only for “industrialized” countries. The paper does have some conclusions that support our starting point in the approach below, namely that, at least in the assessment they use (PISA 2012, all three subjects), there is a tight correlation between markers of absolutely low performance and mean or median performance: “We find a clear and very strong negative correlation between median scores and the incidence of low performance in all three subjects” (p. 33).

All the international assessments have benchmarks for proficiency levels, defined as a number on the cross-country range of IRT scores for the given assessment. Typically, there are some four to seven benchmarks. Some are fairly minimalist and represent the most basic skills. Others are quite high. For PISA Reading, for instance, the benchmark for level 1b (which is one level up from the bottom) refers to:

"Tasks at this level require the reader to locate one or more independent pieces of explicitly stated information; to recognise the main theme or author's purpose in a text about a familiar topic, or to make a simple connection between information in the text and common, everyday knowledge. Typically the required information in the text is prominent and there is little, if any, competing information. The reader is explicitly directed to consider relevant factors in the task and in the text" (p. 162, OECD 2016).

The benchmark for the next-but-highest level is much more demanding:

"Tasks at this level that involve retrieving information require the reader to locate and organise several pieces of deeply embedded information, inferring which information in the text is relevant. Reflective tasks require critical evaluation or hypothesis formulation, drawing on specialised knowledge. Both interpretative and reflective tasks require a full and detailed understanding of a text whose content or form is unfamiliar. For all aspects of reading, tasks at this level typically involve dealing with concepts that are contrary to expectations" (p. 162, OECD 2016).

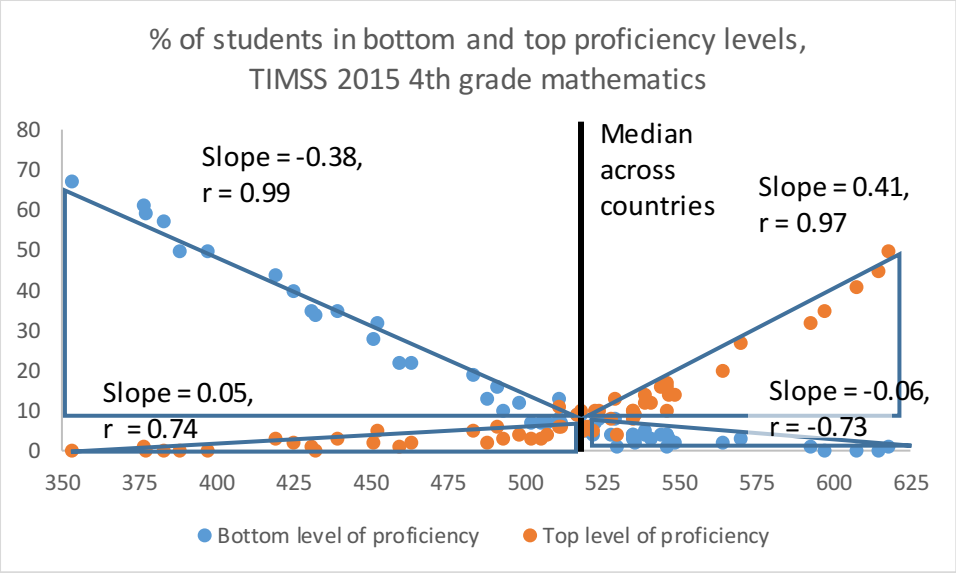
All of the assessments publish something similar.

Each of these benchmarks is associated with a numerical value on the IRT scale. These numerical values are evenly-spaced, at least for the international assessments (it does not seem to be the case for the two regional ones used here), and by construction (if evenly spaced), they are supposed to represent equal increments in the difficulty of the items answered correctly by the students. (In the two regional ones, though the benchmarks are not evenly spaced, the narrative descriptors do suggest roughly cardinal increases in skill.) In that sense, these benchmarks are more than merely ordinal, though one could debate the degree to which the numerical value of the benchmarks is truly cardinal. We address this issue below and it is (inconclusively, at least for economists) discussed in the literature (e.g., Mickewright and Schnepf 2006, and Bruckauf and Chzhen 2016), with most analysts concluding that the measures are more than ordinal but not quite as cardinal as, say, income. For now, we will proceed as if they are cardinal at least in the limited sense that the numerical values of the benchmarks are evenly spaced.

The assessments either publish, or one can obtain, the percentage of children in each country that are at each of the benchmark levels. We can use this, along with the mean performance of each country on the IRT scale, to test the hypothesis that there is a stronger

association between mean scores and the percentage of children above a minimalist benchmark than between mean scores and the percentage of children above a more demanding benchmark. The "slope" of this association can then be used to estimate the "share" of the contribution to the higher score that comes from increasing the percentage of children above a low benchmark, and compare that to the "share" of the contribution that comes from increasing the percentage of children above a high benchmark. Figure 1 illustrates this, using TIMSS 2015 Grade 4 mathematics. Equivalent figures for all of the assessments used in this paper are found in Annex 1.

Figure 1. Relationship between mean performance and percentage of students at two key benchmarks



The country mean scores on the assessment are shown on the horizontal axis. A black line was placed at the median of the country scores. The percentage of children at the low benchmark and the percentage of children at the high benchmark are both shown on the vertical axis. The dots represent the percentages for each of the countries in the sample. Each country is represented twice in the graph: once by a blue marker for the percentage of children at the low benchmark, and once by an orange marker for the percentage of children at the high benchmarks. Finally, two simple, one-variable regressions were carried out between the percentages and the means, and the slope and correlation of the regression are shown. For each of the two benchmarks there are two regressions: one for the countries above the median of the mean country scores, and one for the countries below the median. Thus, there are four slopes and four correlations per assessment.

The figure makes a few things clear.

First, the correlations between either the percentage of children in the low benchmark or at the top benchmark and the mean country scores are high. This can be seen, visually, in the fit of all the slopes, and in the correlation coefficient shown in the graphics. Either percentage is a good predictor of the mean score.

Second, for at least some of the assessments (certainly the international ones), there is a much longer tail to the left of the median than to the right, though this is less the case for the regional assessments, because they generally have a narrower range of variation. Since there is by definition an equal number of countries to the left and to the right of the median, the length of the tail is created by the fact that countries below the median reach "down" into low levels of performance by much more than countries above the median reach "up" into high levels of performance.

Third, the slopes are clearly much steeper for the percentage of children at the low benchmark than for the children above the high benchmark, for countries moving from low to median performance. Using Figure 1, for example, we can see, from the upper left hand triangle, that when countries "move" from the lowest average performance (350 points) to median performance (520 points), the percentage of children at the lowest benchmark drops by around 55 percentage points, creating a steep slope of around -0.38. But the same countries see an increase in the percentage of top students of only about 10 percentage points, for a very weak slope of about 0.05. This relationship between the slopes typically reverses (see Annex 1 for all the cases at once) for countries going from median average performance to high average performance. It may be thought that this is inevitable and is an artefact of the method, but that is not the case, there is no inherent reason why it has to be this way. It is also not due to students in the best-performing countries topping out. In TIMSS 2014 Grade 4 mathematics, for instance, the best-performing country was Singapore. Only 5.1 percent of its children got 100 percent correct (a perfect classical score), yet 50 percent of Singapore's children are at the top TIMSS 2015 Grade 4 mathematics benchmark.

Fourth, a way of summarizing, with a single number, the "contribution" of changing each proportion (reducing the number of children at the low benchmark or increasing the number of children at the high benchmark) suggests itself by the graphic. That method is to calculate the areas of the four triangles and compare them to each other. In particular, the ratio of the area of the high triangle to the left of the median to the area of the low triangle to the left of

the median, and the same ratio for the triangles to the right, can be calculated, and these ratios compared.[10] We do that in Table 2 for all the assessments.

Fifth, note that in most cases the implicit regression lines cross each other at the median of the mean levels of country performance. In most cases, the cross-country median proportion of children performing at the low benchmark was roughly equivalent to the cross-country median proportion of children performing at the high benchmark, so the comparison of the slopes makes sense. The fact that the implicit regression lines, in most cases, cross each other at the median level of mean country performance suggests approximate numerical equivalence of proportions of children at the chosen benchmarks, and that we are not "cheating" by making the cut-offs too high or too low. In some cases, the median proportion of children performing at the high benchmark (or low) was so small compared to the median proportion of children performing at the low (or high) benchmark that the slopes would likely be different by construction.[11] In those cases, to be more conservative, we moved the benchmark and took, say, the top two benchmarks or the bottom two. In the case of PISA and SACMEQ, we generally took both the bottom two and the top two, because they have so many benchmark levels that the data were a bit jumpy.

Lastly, note that though the graphic graphically (to be redundant) conveys symmetry (around the median—that is symmetry around the vertical axis), the point is quite the opposite: the relationship is very non-symmetrical: to the left of the global median, the percentage of children achieving at the low benchmark is highly *negatively* associated with increases in countries' means, and the percentage of children achieving at the high benchmark is only weakly positively associated with increases in countries' means. To the right of the global median, the opposite holds: the percentage of children achieving at the low benchmark is very weakly associated with increases in the countries' means, and the percentage of children achieving at the higher benchmarks is strongly associated with increases in countries' means.

[10] In reality since the bases are the same, the ratio of the areas is the same as the ratio of the slopes.
[11] SACMEQ mathematics, in particular, appears rather optimistic about the higher benchmarks, so we had to take the top four benchmark levels to make those more or less equivalent in size the lowest two levels.

Table 2. Comparison of slopes for all assessments					
Assessment	Ratio top slope to bottom slope, left of median	Ratio bottom slope to top slope, right of median	Assessment	Ratio top slope to bottom slope, left of median	Ratio bottom slope to top slope, right of median
PISA 2015 Reading	6.1***	0.54***	PIRLS 2011	7.5***	0.35***
PISA 2015 Mathematics	6.0	0.38***	SACMEQ III Reading	2.2	0.32*
PISA 2015 Science	6.0***	0.48***	SACMEQ III Mathematics	6.9***	0.66***
TIMSS 2015 Grade 4 Mathematics	8.2***	0.15*	TERCE Grade 3 Reading	4.0**	1.0***
TIMSS 2015 Grade 8 Mathematics	9.8***	0.26***	TERCE Grade 3 Mathematics	10.0**	2.1**
TIMSS 2015 Grade 4 Science	12.3***	0.21***	TERCE Grade 6 Reading	1.9*	0.72*
TIMSS 2015 Grade 8 Science	7.5***	0.49***	TERCE Grade 6 Mathematics	2.7***	0.82***
For a test of the difference between slopes: * p value less than or equal to 0.05 ** p value less than or equal to 0.01 *** p value less than or equal to 0.001					

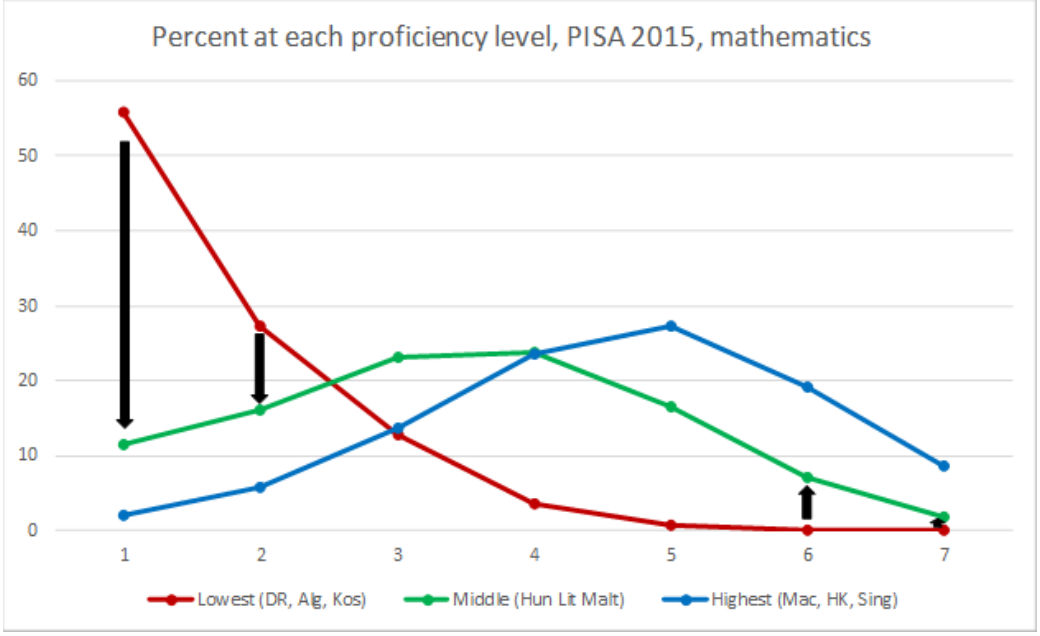
It should be noted that these assessments cover a wide range of countries, with low-income countries being largely unrepresented.[12]

[12] The breakdown is as follows. PISA 2015: 0.63 HI, 0.30 UMI, 0.07 LMI. TIMSS 2015 Grade 8: 0.72 HI, 0.22 UMI, 0.07 LMI. TIMSS 2015 Grade 4 Mathematics: 0.84 HI, 0.11 UMI, 0.05 LMI. TIMSS 2015 Grade 4 Science: 0.85 HI, 0.09 UMI, 0.06 LMI. PIRLS 2011: 0.78 HI, 0.16 UMI, 0.07 LMI. TERCE: 0.40 UMI, 0.53 LMI, 0.07 LI. SACMEQ: 0.27 UMI, 0.13 LMI, 0.60 LI.

The table also shows that the differences between the top and bottom slopes (on either side of the median) are generally highly statistically significant, aside from the obvious high substantive significance. In terms of substantive significance, across all the assessments, the top triangle on the left of the median (changes in the percentage of children at the low benchmark in going up to the median in mean test scores) is much larger than the bottom triangle on the left of the median (changes in the percentage of children at the high benchmark in going up to the median in mean test scores). So, the reduction in the percentage of children at the low benchmark contributes 6.9 times more on average, than an increase in the percentage of children at the high benchmark, in the transition from low mean-scoring countries to median mean-scoring countries. On the right of the median, the situation is reversed: the bottom triangle (changes in the percentage of children at the low benchmark) is only 0.62 (on average) as large as the top triangle. Reductions in the percentage of children at the low benchmark still count in going from median to high mean country performance, but not as much as the increase in the number of children at the high benchmark.

In previous writing by one of the authors (see Crouch and Rolleston 2017), data similar to that used above and in Annex 1 have been portrayed slightly differently. For instance, the data on TIMSS 2015 Grade 4 mathematics, showing that as countries move from low to middle performance, the drop in the percentage of students at the lowest benchmark is much higher than the increase at the high benchmark, is made via Figure 2. This graphic takes the percentage of children at *all* PISA benchmark levels (benchmark levels are on the horizontal axis), for the countries with the top three mean PISA scores, the countries with the three middle scores, and the countries with the three bottom scores. This graphic may be more comprehensible than Figure 1, but it conveys less information. We present the graphic for two reasons. First, to enable a link to the companion RISE publication by Crouch and Rolleston (2017), and because the micro data analysis to follow essentially applies this graphic to the micro data, as can be seen via a comparison of Figure 2 here and Figure 5 and Figure 6.

Figure 2. "Histograms" of PISA 2015 mathematics results



3.2 Evidence from micro data

3.2.1 The data

This section makes use of data from two international programmes which test the cognitive skills of pupils at the primary school level: TIMSS (Trends in International Mathematics and Science Study) and SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality). TIMSS is a large programme spanning all continents with a history going back to 1964. It tests at both the primary and lower secondary levels, specifically grades four and eight (the grade four data were used). SACMEQ is a relatively new and regional programme focussing just on the primary level, specifically Grade 6. It first ran tests in 1995. The analysis that follows uses data from sixty-four TIMSS countries across four years (2003, 2007, 2011 and 2015) and fifteen SACMEQ countries across two years (2000 and 2007). One country, Botswana, is found in both programmes.

One important limitation of the TIMSS data in terms of our purposes must be made clear. Developing countries included in the data we use are to a large extent in one world region: Middle East and North Africa (to use the region terminology of the World Bank). Specifically, seven non-high income countries are from this region, of which six have data for more than one year, compared to four countries in Latin America (including Chile, now a high-income

country), one in Sub-Saharan Africa (Botswana)[13] and four in East Asia and Pacific. Crucially, only Chile from these last three regions has more than one year's data. In Europe and Central Asia four developing countries, namely Turkey, Kazakhstan, Armenia and Georgia, all have more than one year of data, whilst Moldova and Azerbaijan have just one year.

Only mathematics test data were used, not data for any other subject (science in TIMSS and reading in SACMEQ).

3.2.2 The classical scores behind the IRT scores

Widely quoted TIMSS and SACMEQ scores are derived from statistical processes that use item response theory (IRT). Each of these programmes apply a final and simple rescaling step to their IRT scores which in itself is not an IRT process. The rescaling is determined in a base year, and makes the mean for all pupils across all countries in the base year 500, and the standard deviation 100. Clearly 500 in TIMSS represents a higher level of achievement than 500 in SACMEQ, given that the great majority of TIMSS countries in the base year were developed countries. The TIMSS and SACMEQ base years were 1995 and 2000.

Underlying the IRT-generated scores are the classical scores ("percent correct") which most people would easily comprehend. This section explores the relationship between the two types of scores largely as a prelude to a subsequent examination of how the choice of score type influences the measurement of inequality.

The TIMSS and SACMEQ tests consist of items (questions), each of which carry a maximum score, meaning any pupil's maximum possible overall score would be the sum of the maxima across all items. The overall score can of course be expressed as a percentage correct statistic. These are the classical scores. IRT-generated scores are essentially superior pupil-level overall scores produced from the raw test data. They are superior in the sense that they are better at differentiating the performance of pupils, because they take into account the statistically-derived level of difficulty of each item each pupil answers correctly. Two pupils with the same classical score are thus very likely to have different IRT scores, depending on which items they answered correctly.

In multi-year programmes such as TIMSS and SACMEQ, IRT scores moreover include an adjustment in one year to make scores comparable to those in a previous year. This is

[13] South Africa is not included as South Africa's only participation in TIMSS grade 4, in 2015, was through an easier test known as the 'Numeracy assessment'. Given the focus in the current paper on comparing IRT and classical scores, it was necessary to use data from tests at the same level.

achieved through common, or anchor items repeated in the different years. The presence of anchor items obviously makes the security, or secrecy, of the tests, or at least anchor items, enormously important. After being administered, all test booklets must be removed from testing venues.

A further complexity, found in TIMSS, is matrix-sampling, or the use of different test booklets for different pupils, with common items spanning these booklets. This has the benefit of allowing for the testing of a wider variety of items, though the cost is statistical complexity as IRT scores must be based in part on imputations of what pupils would achieve across items they did not have access to.

Finally, IRT scores in TIMSS are further fine-tuned through the use of data from the TIMSS subject other than mathematics, namely science. For instance, a high score in science can inflate the pupil's mathematics score. Even background information relating to, for example, the pupil's home background is used to a limited degree to verify and possibly adjust scores.[14]

A part of the interest in gauging the inequality of test scores, or the cognitive skills they represent, arises out of the interest of economists in comparing educational inequalities to inequalities in other areas, in particular income (Sahn and Younger, 2007). Classical scores enjoy an intuitive appeal for some economists largely due to the fact that their minimum is zero and that their construction is clearly a true metric: answering 90 percent correct is as much "more correct" over 80 percent than 70 percent is over 60 percent. However, to educators and psychometricians this is unsatisfactory, as which items students answered correctly matters. A student scoring 90 percent correct, but who was wrong on the most difficult items is less capable than one who scored 90 percent correct, but got quite a few of the most difficult items correct. Percent correct also clearly has an absolute minimum. IRT-generated scores have no absolute minimum. Thus, classical scores appear to lend themselves to inequality statistics, such as the Gini coefficient, mostly used to gauge income inequality. Whether this is true is explored below.

Exploring the classical scores behind the IRT-generated scores of programmes such as TIMSS and SACMEQ, something which is seldom done, can serve purposes beyond assisting economists measuring educational inequality. The exercise can assist in explaining the meaning of IRT-generated scores in different programmes, a matter of importance given that many education analysts and economists do not understand IRT scores very well, whilst they do understand classical scores.

[14] Foy and Yin, 2016: 13.20.

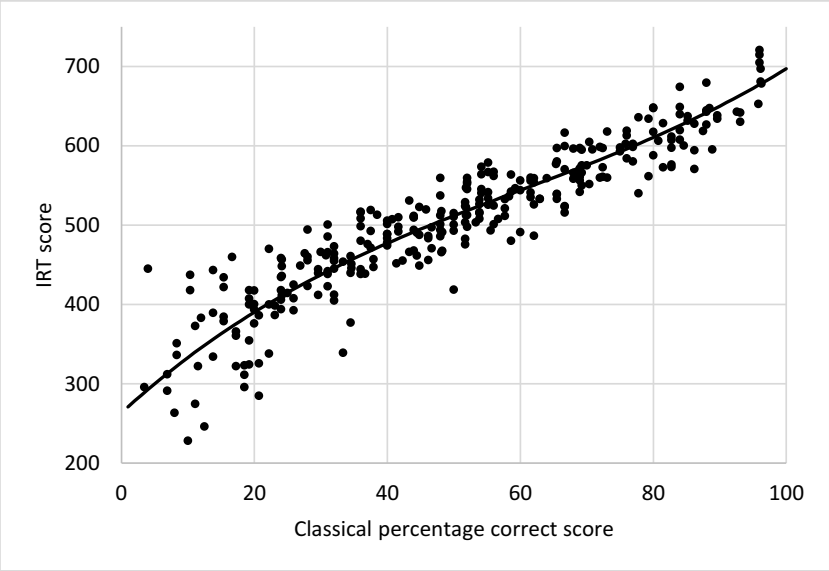
Classical scores, as in a score of 30 out of a total of 50, are not provided in the TIMSS data (they are in the SACMEQ data), and must thus be calculated from the item-level data. To illustrate how this was done, the process for the 2015 data is briefly described. The publicly available 2015 TIMSS data include multiple-choice responses provided by pupils, and scores assigned to remaining non-multiple-choice questions, referred to as 'constructed response questions'. The TIMSS metadata include maximum scores possible per constructed response item, and what the correct multiple-choice responses are. Each pupil responded to one of fourteen sets of mathematics items. These sets, or 'booklets', shared many items. Each set contained around twenty-six items, half of which were multiple choice items, each counting for one mark. A few of the constructed response items counted for more than one mark, giving a total possible score of around 30 per pupil. The scores obtained per pupil were calculated from the data, and these were divided by the maximum possible per booklet, which could differ across the fourteen booklets. (Each of the fourteen booklets also contained science questions, but these were ignored for this analysis.) The result was the percentage correct score per pupil used to generate Figure 3 below.

The IRT scores on the vertical axis were calculated by finding the mean, per pupil, across the five 'plausible values' in the TIMSS data. There is not one IRT score per pupil, but five, which differ from each other, for the same pupil, by an average of 31 TIMSS points, or about 0.31 of a global standard deviation across pupils. The fact that there are five values per pupil reflects the uncertainty around the determination of each pupil's fundamental ability, given for instance the fact that each pupil is asked to respond only to a sample of all the TIMSS items used in the year. The 31 TIMSS points difference is large, yet the mean plausible value per pupil produces country-level means which are identical to official country means in the case of forty of the forty-five countries of 2015. For five countries, all relatively poorly performing countries, the calculated mean minus the official mean is non-zero, in the range of -8 to 2 TIMSS points, with the most extreme being Indonesia (-8) and Iran (-7). The plausible values exist primarily to allow analysts to calculate aggregate statistics, with confidence intervals, using tools produced by TIMSS. Their main purpose is not to produce IRT scores per pupil. This explains why one would not expect a perfect match between country means derived from pupil-level means (as calculated here), and official country means.

A key concern when classical scores are analysed, is floor and ceiling effects, or the extent to which pupils are concentrated close to the minimum (in the case of a difficult test) or the maximum (in the case of an easy test). Overall, only 0.7 percent of TIMSS Grade 4 pupils in 2015 reached the classical ceiling, meaning all their responses were correct. Only four

TIMSS countries saw over 3.0 percent of pupils reach this ceiling: Taiwan, Hong Kong, South Korea and Singapore (the latter had the highest figure, of 5.1 percent). At the other extreme, 0.3 percent of pupils overall were at the zero percent correct floor, with Kuwait displaying the highest value at 3.1 percent. These ceiling and floor statistics obtained from our analysis agreed with statistics reported by TIMSS analysts and shared with us, a fact that strengthened our conclusion that we had processed the TIMSS item data correctly. Floor and ceiling effects should be less of a concern when IRT scores are analysed, given that, for instance, the floor is diffuse in the sense that pupils with a zero classical score could have a variety of IRT scores, based in particular on their performance in science (in the case of TIMSS).

Figure 3: Classical and IRT scores in 2015 TIMSS Grade 4 mathematics



Source: TIMSS 2015 grade 4 mathematics microdata downloaded at https://timssandpirls.bc.edu.

Note: Points represent 350 randomly selected pupils, from 280,805 pupils in the TIMSS database. The regression trendline, which is cubic, is calculated using all the data.

The relationship between the two types of scores seen in Figure 3 has two features, which suggest that measures of inequality will be inflated for worse performing countries if one uses IRT scores, relative to classical scores. The first feature is an inverted S relationship. This can be seen in the trendline, which represents a cubic function. The slope is clearly flatter in the classical score range of 40 to 80, relative to the adjoining ranges. The second feature is more vertical dispersion around the trendline towards the left-hand side of the graph. The interquartile range in IRT scores is 60 TIMSS points for pupils with a score of

around 20 percent correct, relative to 35 TIMSS points for those with a score of around 80 percent correct.

These patterns place in doubt the assumption used, for instance, by Sahn and Younger (2007: 1) that IRT scores can be straightforwardly considered cardinal values for the purposes of measuring educational inequality. This is an addition to the general dubiousness around the cardinality of test scores, whether classical or IRT-generated. To illustrate, can we really say that the gap, in terms of human capital, between Person A with percentage correct of 40 and Person B with 50, is the same as the gap between Person B and Person C with 60, considering in particular that a higher score means having answered more difficult, and hence more 'valuable' questions correctly? Clearly, there is far less meaning in such assertions than in similar assertions relating to monetary income or wealth.

If one runs correlations between IRT and classical scores for the 157 existing combinations of countries and years (64 countries[15] across four years, with many countries having fewer than four years), correlation coefficients range from a high of 0.95 for Kazakhstan 2007 to 0.68 for Yemen 2007, the median being 0.92 for Sweden 2011.

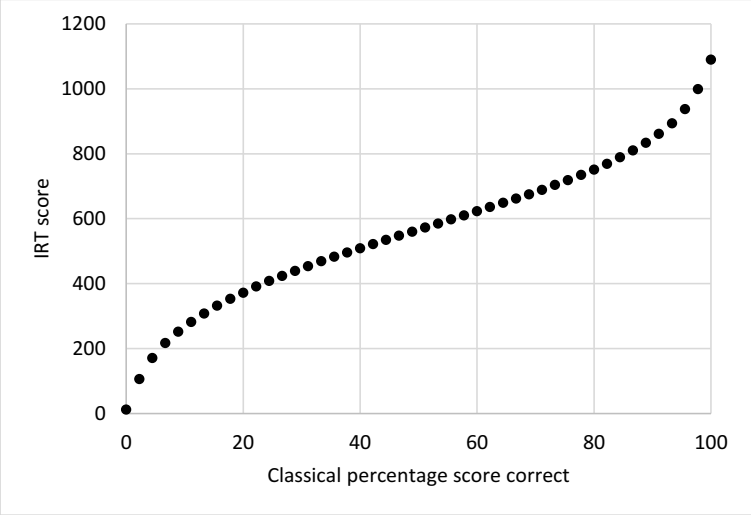
A graphic similar to Figure 3 for TIMSS was produced for SACMEQ III mathematics in Figure 4. It is clear from Figure 4 that the non-linear inverted S of TIMSS reappears in SACMEQ. In this sense, the SACMEQ IRT scores are typical and the relative "lack of cardinality" issues would affect SACMEQ as well. The implications of this for measuring inequality are discussed below.

Note that relatively few SACMEQ pupils display floor or ceiling values: in 2007, 0.1 percent of pupils were at the floor (with Zambia having the highest value of 0.4 percent), and only four pupils overall at the classical ceiling of 100 percent.[16]

[15] What TIMSS refers to as 'benchmarking participants', mainly regions within countries, were excluded for all analyses presented here.

[16] Reproducing Figure 3 using SACMEQ 2007 data, the most recent SACMEQ data to which we had access, results in the curious patterns of Figure 4. For thirteen (including Zanzibar as a "country") of the fifteen participating countries one-to-one relationships exist between the classical score and a score referred to as a 'logit score' within the data, and a 'Rasch score' in the available documentation, meaning essentially an IRT-derived score. This relationship emerges whether one uses the original IRT pupil score in the data, or the rescaled IRT score producing a mean of 500 and a standard deviation of 100 (it is the latter which is used for the graph). Two countries, Mozambique and Tanzania, each display one-to-one relationships which are very close to those of the thirteen other countries, though not identical. It is probably not a coincidence that these two countries are two of the three countries which did not use mathematics tests which were in English (the third was Zanzibar). With regard to the classical scores, these appeared in a distinct variable in the dataset , but to verify them it was established that they could be recalculated using item-level responses in the dataset. The IRT scores, of which just one appeared per pupil (there were thus not various 'plausible values' per pupil, as in TIMSS), produced country-level weighted means which corresponded perfectly with

Figure 4: Classical and IRT scores in 2007 SACMEQ Grade 6 mathematics



Note: The points in the graph represent the one-to-one relationship of the thirteen countries discussed in the text. All these countries display exactly the same one-to-one relationship between classical scores and IRT scores.

---

country means published by UNESCO's International Institute for Educational Planning (IIEP), a key organisation working on the technical aspects of SACMEQ in 2000 and 2007.

The one-to-one relationship seen in Figure 4 is curious as IRT scores should, by design, differentiate between pupils with the same classical scores. Each classical score should thus translate to various IRT scores, and this is what one finds in TIMSS, as seen in Figure 3. (Importantly, the vertical dispersion seen in Figure 3 would also be substantial if the graph represented the results from just one of the fourteen TIMSS booklets.) For an exercise unrelated to the current paper, one of the authors calculated IRT scores, or 'person locations', using RUMM (the IRT software employed by the IIEP) and the 2007 original item-level responses, for South Africa only. This produced 725 unique IRT score values, against the 45 unique values seen in the official mathematics dataset.

There is not enough information in the available documentation, for the 2000 or 2007 SACMEQ waves, to explain the one-to-one relationship between classical and IRT scores in the SACMEQ data. It should be emphasised that SACMEQ is far less transparent and technically developed as a programme than TIMSS. This is something one would expect given SACMEQ's relatively short history and low level of resourcing. However, it is noteworthy that even relative to another regional and developing country programme, which like SACMEQ has been historically linked to UNESCO, namely Latin America's LLECE, SACMEQ compares poorly. Whilst LLECE allows for the downloading of data and basic technical documents off a web-based portal, SACMEQ does not. Special offline requests are needed to access SACMEQ data, and technical documents have been difficult to access even amongst those with access to the data.

It should be underlined that despite the anomalies pointed out here, and problems around the availability of technical documentation, SACMEQ's role in the regions of Southern Africa and East Africa, in terms of alerting policymakers to specific educational quality and equality problems, and in terms of exposing analysts to vital education data, has been immense. SACMEQ already offers considerable value to research and policymaking, though with technical improvements, it could offer even greater value.

The 'one-to-one problem' seen in the 2007 SACMEQ mathematics data was also seen in the 2000 data, and for both years the anomaly also appeared in the reading test data.

3.2.3 On the distribution of improvements: cross-sectional micro evidence

The evidence presented in section 3.1 (and more massively in Annex 1) is reproduced in this section, but this time using micro data from TIMSS 2015 Grade 4 mathematics, and from SACMEQ mathematics. The conceptual "bridge" between the two sections is Figure 2 (and the similar figures used in Crouch and Rolleston 2017). For each of the two assessments, we show the distribution of students for low-performing, middle-performing, and high-performing countries. In addition, the data are shown using both IRT scores and classical scores ("percent correct") so as to explore whether conclusions differ significantly depending on an intuitively cardinal measure, the classical scores, and one which, one could argue, is not so obviously cardinal.

In these figures, the vertical bars represent the TIMSS (or SACMEQ) benchmarks. The letters refer to the areas between the curves. Area A, therefore, represents the contribution made by shifts in the proportion of students at the low level within countries to movement between low country means and middle country means, area B the contribution made by shifts in the proportion of students at the high level within countries, to movement of countries from low to higher average levels, and so on.

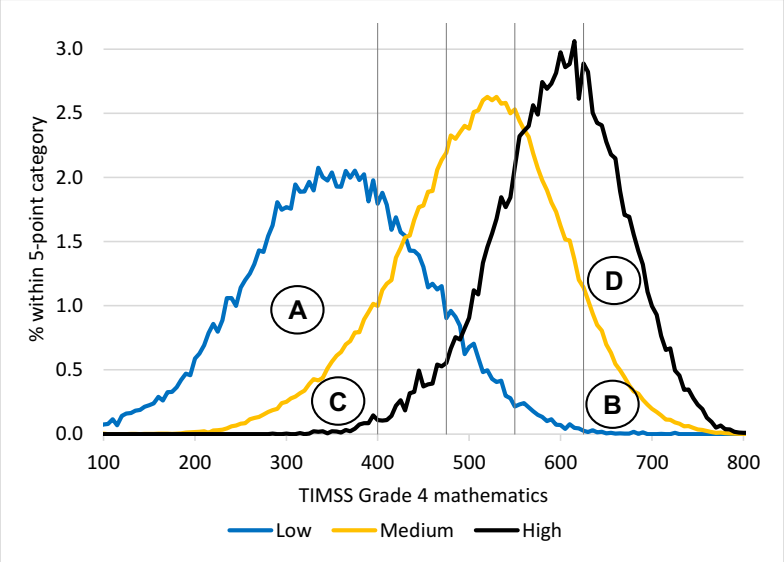
The TIMSS benchmarks for the four bars, and for the selection of countries,[17] is the following:

625 Advanced
550 High
475 Intermediate
400 Low

Note that they are all spaced 75 points apart.

[17] A 'low' country in the above graph (and below) is one with a TIMSS mean below 400 (the official minimum threshold for qualifying as 'Low'). A 'Medium' country has a TIMSS mean 400 or above, but less than 550 (550 official minimum for 'High') and the 'High' here is 550 or above.

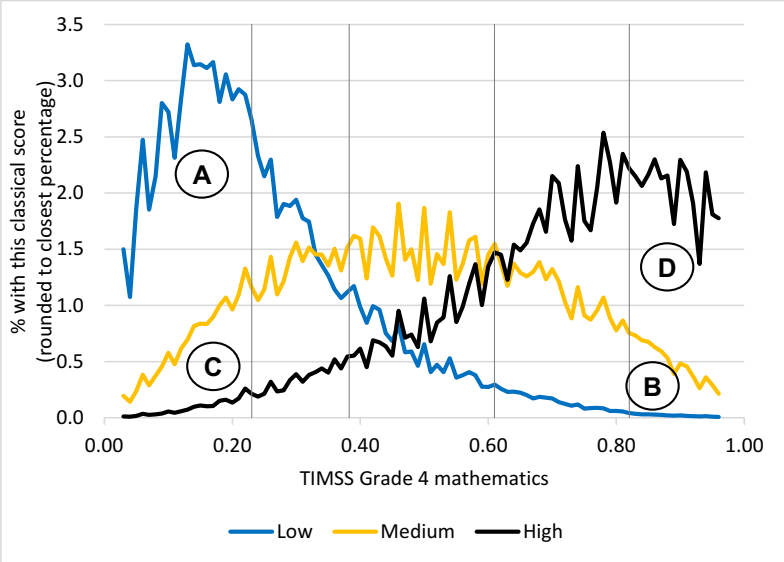
Figure 5: Distribution of TIMSS cognitive skills using IRT scores



Source: TIMSS Grade 4 mathematics microdata of 2015, 2011, 2007 and 2003.

Note: Vertical lines represent the cut-offs for the international benchmarks. The curves represent 11 (low), 47 (medium) and 6 (high) countries, ranked according to country mean. Only the most recent results for each of the 64 countries were used. Within countries TIMSS weights are used, but averages across countries are not weighted. Horizontal intervals of 5 TIMSS points are used, with each pupil's score being rounded down to a multiple of 5.

Figure 6: Distribution of TIMSS cognitive skills using classical scores



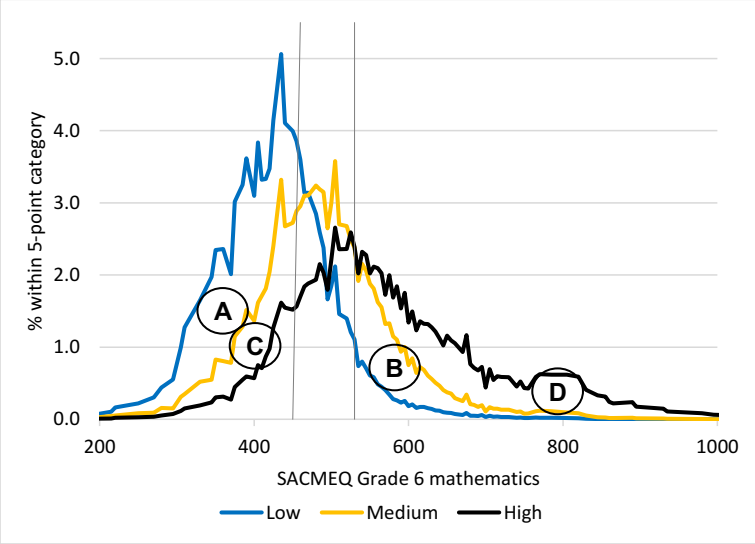
Source: TIMSS Grade 4 mathematics microdata of 2015, 2011, 2007 and 2003.

Note: Vertical lines represent classical score equivalents of the international benchmarks. Notes on the previous graph which could apply here, apply here too. Moving average smoothing across five percentage points is used here.

Table 3: Gains at the bottom and top ends		
	Low to Medium	Medium to High
Using TIMSS IRT scores		
Area between two curves below 400	-54 (A)	-13 (C)
Area between two curves 625 and above	10 (B)	27 (D)
Using TIMSS classical scores (Figure 6)		
Area between two curves below 400	-43 (A)	-13 (C)
Area between two curves 625 and above	7 (B)	27 (D)
Using SACMEQ IRT scores (Figure 7)		
Area between two curves below 460	-30 (A)	-16 (C)
Area between two curves 530 and above	22(B)	27 (D)
Using SACMEQ classical scores (Figure 8)		
Area between two curves below 460	-29 (A)	-13 (C)
Area between two curves 530 and above	23(B)	27 (D)
Note: Values refer to the percentage of all pupils.		

Whether one uses classical scores (more intuitively cardinal) or IRT scores (less intuitively cardinal) does not make much difference. For TIMSS, the areas in moving from middle to high are exactly the same (C and D, rows 1, 2, 3, and 4) in moving from low to middle are fairly close (A and B, rows 1, 2, 3, and 4). For SACMEQ they are the same for any practical purpose. Furthermore, these findings replicate directionally and within reasonable magnitude, the more macro findings shown in Table 2.

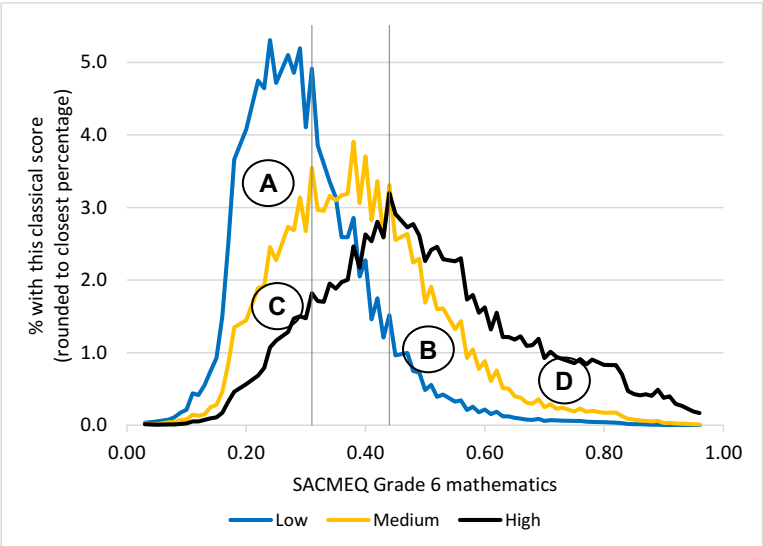
Figure 7: Distribution of SACMEQ cognitive skills using IRT scores



Source: SACMEQ Grade 6 mathematics microdata of 2007 and 2000.

Note: Vertical lines represent the cut-offs for the official SACMEQ levels of 'Basic numeracy' (460) and 'Beginning numeracy' (530). The curves represent 3 (low), 8 (medium) and 4 (high) countries, ranked according to country mean. Results were pooled across the two years per country. Within countries SACMEQ weights are used, but averages across countries are not weighted. Horizontal intervals of 5 SACMEQ points are used, with each pupil's score being rounded down to a multiple of 5. Moreover, smoothing was achieved through a moving average covering five data points (so 25 SACMEQ points).

Figure 8: Distribution of SACMEQ cognitive skills using classical scores



Note: Vertical lines represent classical score equivalents of the official SACMEQ benchmarks. Notes on the previous graph which could apply here, apply here too. Moving average smoothing across five percentage points is used here.

3.2.4 Time-based evidence on the distribution of learning outcomes

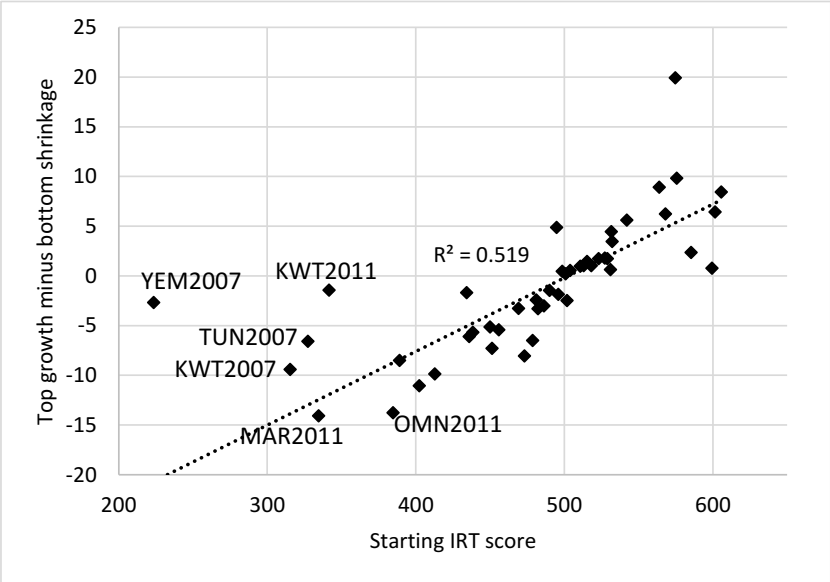
3.2.4.1 Changes in proportions at different benchmarks over time (changes in "cognitive poverty")

The cross-sectional pictures provided above suggest that as the educational quality of low-performing countries improves, there should be larger reductions in weak performance than increases in high performance. This pattern is in fact what one finds if one analyzes actual country-specific changes over time. In the TIMSS Grade 4 data from the four years, there were sixty-one instances of significant change over time represented by two consecutive points in time for the same country. One country could be represented by more than one instance, for example, Qatar saw a significant improvement between 2007 and 2011, and again between 2011 and 2015, creating two instances for Qatar. An average annual gain of 1.5 TIMSS points was used as a threshold for considering a change significant, which would be in line with official TIMSS reports.[18] Of the sixty-one instances, forty-eight involved significant improvements (as opposed to significant declines) and are plotted in Figure 9 below. Of twenty-six instances of improvers where the starting point was less than 500 TIMSS points, all but two involved more shrinkage in the number of pupils below the low TIMSS benchmark (400) than growth in the number of pupils reaching the advanced benchmark (625). The two exceptions represent two rich countries: Austria and Norway.

A relevant question would be whether any countries improved their average scores by only reducing the proportion of children below the lowest benchmark (400 TIMSS points), without at the same time increasing the proportion above the highest benchmark (625). There were in fact four such instances, all from the Middle East and North Africa region: Iran in 2003 to 2007, Kuwait in 2007 to 2011, and Morocco and Bahrain, both in 2011 to 2015.

[18] Mullis, Martin, Foy and Hooper, 2016: Exhibit for Grade 4 titled 'Differences in Mathematics Achievement Across Assessment Years'.

Figure 9: TIMSS shrinkage at the bottom versus growth at the top



Note: The graph draws from the data of 34 TIMSS countries, and 61 instances of significant change. The vertical axis is the increase in the percentage of pupils at or above the advanced benchmark (625) minus the decrease in the percentage of pupils below the low benchmark (400). Thus a negative value means a decrease at the bottom which exceeds the increase at the top. To illustrate the labelling, 'TUN2007' refers to change between Tunisia in 2007 and the next TIMSS year for Tunisia, which would be 2011.

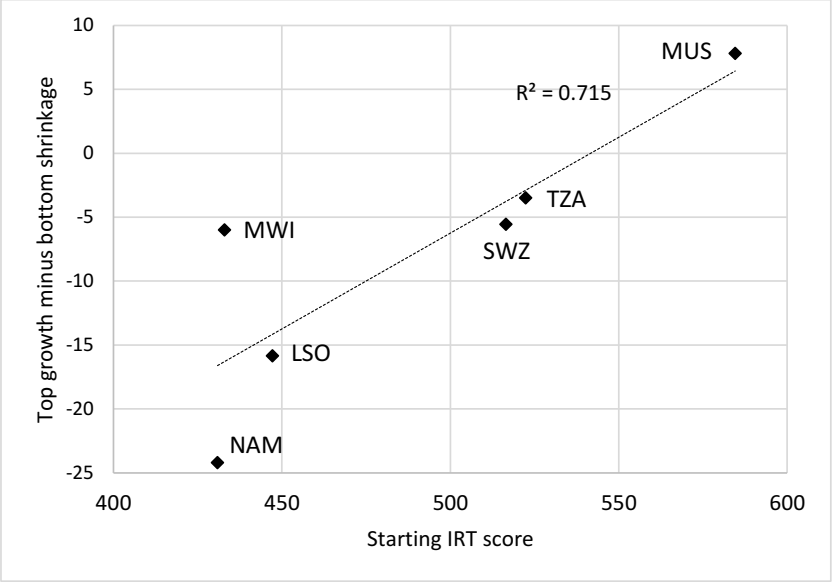
Figure 9 adds to an analysis by Mullis *et al* (2016: 58), who examine the improvements amongst TIMSS Grade 4 countries, between 1995 (but in some cases 2003) and 2015, focussing on improvements at the 10th and 90th percentiles. They conclude that national gains are driven more by the desired change at the bottom end of the performance spectrum than the top. Of eighteen countries, all but four saw larger, often much larger, improvements at the 10th than the 90th percentile. The present analysis, by including more developing countries, establishes that the movement is towards less "cognitive poverty."

Just six SACMEQ countries were considered to have made significant improvements in their national mathematics score between 2000 and 2007.[19] These countries are illustrated in Figure 10 below, which follows the approach of the previous TIMSS graph. Generally, the six SACMEQ countries did see larger reductions at the bottom than gains at the top. The exception is Mauritius (MUS), by far the best performer of all fifteen SACMEQ countries in both 2000 and 2007. The thresholds used to define the bottom and the top for the purposes

[19] Makuwa, 2010.

of this graph were the 460 and 645 SACMEQ scores, minimum values for the official SACMEQ levels 'basic numeracy' and 'mathematically skilled'.

Figure 10: SACMEQ shrinkage at the bottom versus growth at the top



One matter not taken into account in any of the above analysis is changes in enrolment ratios over time. These can be large in developing countries, and were they to be taken into account they would change the analysis a bit, but almost certainly not to the degree that the patterns would substantially change. The necessary adjustments are not easy to undertake, largely due to data problems relating to enrolment ratios, as explained in Gustafsson (2015). Had enrolment ratios been taken into account, one change would have been the inclusion of Mozambique in Figure 10. The case of Mozambique is particularly interesting. Mozambique's SACMEQ means declined between 2000 and 2007. This was driven by large growth at lower levels of performance due to a substantial expansion in access to primary schools. However, there was also growth in the number of pupils at middle and higher levels of performance. The latter phenomenon was largely missed as it was obscured by the decline in the mean (Taylor and Spaull, 2015).

3.2.4.2. Inequality over time

The Sustainable Development Goals mark an important shift towards a stronger emphasis on measuring and tackling inequalities, including educational inequalities. Whilst there has been some analysis of whether schooling systems are on average becoming better, with respect to the cognitive skills of pupils, research into whether these skills become more

equitably distributed is still limited – key existing studies are those of Freeman, Machin and Viarengo (2011) and Oppedisano and Turati (2015), noted in the context of Table 1 above.

The analysis that follows focusses on whether the TIMSS and SACMEQ data, which cover a maximum of four and two time periods respectively, point to a global trend towards greater or less within-country inequality with respect to test scores. Four commonly used measures of inequality are used. The analysis also considers the relationship between changes in inequality, on the one hand, and the direction of change in the mean score and the level of development of the country, on the other. The impact of using classical test scores, as opposed to IRT scores in the analysis is also covered.

Table 4 below focuses on the consistency of measures of inequality when one uses just IRT scores, or when one uses just classical scores. Each of the ten TIMSS values in the table is the correlation coefficient across two variables with ninety-three observations, where each observation represents an earlier and a later value for a country. The sixty-one instances of change discussed in relation to Figure 9 above are a sub-set of the ninety-three observations, with the latter set including even instances of negligible change in the mean score. If one compares country-level IRT-based means across years, for instance Qatar in 2007 and Qatar in 2011, one obtains a high correlation coefficient of 0.96. A similarly high correlation of 0.96 is found if one bases one's country means on classical scores. The correlation across country means are presented mainly as a benchmark against which to view correlations across the country-level inequality values presented in the next four rows of the table.

Table 4: Correlations across time periods

	TIMSS (n = 93)		SACMEQ (n = 14)	
	IRT	Classical	IRT	Classical
Mean	.96	.96	.89	.89
Theil T	.93	.96	.87	.82
Generalised entropy index with parameter - 1.0	.84	.91	.33	.82
Gini coefficient	.96	.96	.87	.82
90^{th} percentile / 10^{th} percentile	.93	.91	.84	.78
Note: Correlation coefficients presented here and in the following table use pupil weights.				

If one focuses on TIMSS, the correlation, at the country level, between one year's level of inequality and another's year's level of inequality, using the same type of score (IRT or classical), is high. This could suggest that robust measures of inequality are obtained, regardless of the type of score. However, as will be seen below, the type of score used does influence conclusions around inequality in substantive ways. Turning to SACMEQ, mostly high correlations are found, though they are lower than the TIMSS correlations. This would be consistent with the hypothesis that SACMEQ, clearly a less technically developed testing programme than TIMSS, is less able to gauge accurately the performance of pupils or countries, and levels of within-country inequality.

Table 5 turns to the matter of correlations between IRT-based values and classical-based values. To illustrate the approach, in the case of TIMSS the second column reflects correlations along the ninety-three observations discussed above, and the aim is to examine the robustness of measures of country-level change between one year and another. Thus one observation could have, in one variable, the change in Qatar's Gini coefficient between 2007 and 2011, using IRT scores and, in the second variable, Qatar's 2007 to 2011 Gini change using classical scores. The first column in the table provides correlation coefficients along 157 observations, where each observation represents a country in a year (so Qatar 2007, 2011 and 2015 would occupy three observations), and where the two variables would contain the same statistic with one using IRT scores and the other classical scores.

The correlations across country means, using the two score types, is high, 0.98 for TIMSS and 1.00 for SACMEQ. The high correlation for SACMEQ should not come as a surprise considering the one-to-one relationship discussed earlier and seen in Figure 7. In the case of SACMEQ, even if country rankings in terms of means may not be extremely consistent from one year to another (last column of Table 4), the IRT and classical scores are so correlated to each other that on the whole it does not matter for the rankings which is used when gauging within-country inequality, or just country means (the last two columns of Table 5).

The situation is very different in TIMSS. Here, although for gauging changes in the mean the choice of score type does not play a very large role (second column of Table 5), which score type one chooses has large implications for gauging changes in the level of inequality. In fact, two of the four correlation coefficients dealing with inequality in the second column are negative, meaning a move towards less inequality using one score type appears as a move towards more inequality using the other. An obvious question in terms of the concerns of this paper is which of the two score types is better for measuring inequality.

Table 5: Correlations between classical and IRT scores

	TIMSS		SACMEQ	
	Corr. across country-year values (n = 157)	Corr. across change values (n = 93)	Corr. across country-year values (n = 29)	Corr. across change values (n = 14)
Mean	.98	.91	1.00	1.00
Theil T	.82	.04	.92	.97
Generalised entropy index with parameter -1.0	.67	-.19	.49	.50
Gini coefficient	.87	.30	.94	.98
90^{th} percentile / 10^{th} percentile	.84	-.14	.96	.99

Figure 11 below depicts the relationship between Theil T[20] measures of inequality based on classical scores and such measures based on IRT scores, thus elaborating on what was seen in Table 5. The relationship in SACMEQ works very differently to the relationship in TIMSS. For the same classical-based Theil T value, TIMSS converts to a lower IRT-based value than SACMEQ. Comparisons across different measures of inequality for the same country where the score type or programme differ appear not to be meaningful. This can be seen in relation to Botswana (BWA).

If one focusses on SACMEQ, with its peculiar one-to-one problem, one finds that the incomparability of individual Theil T values based on different score types is attributable to two things: the fact that zero in the classical system is not equal to zero in the IRT system, and the non-linear relationship between the two types of scores (seen previously in Figure 4). If one forced the equivalence of the two zeros, by saying that IRT scores were simply the classical scores multiplied by some positive factor, then all four measures of inequality would be identical, regardless of score type. This is due to what is known as the scale invariance of the four measures. However, as soon as one departs from the equivalence of the two zeros, inequality measures change when the score type changes. To illustrate, we could calculate the IRT score *I* for pupil *p* as follows:

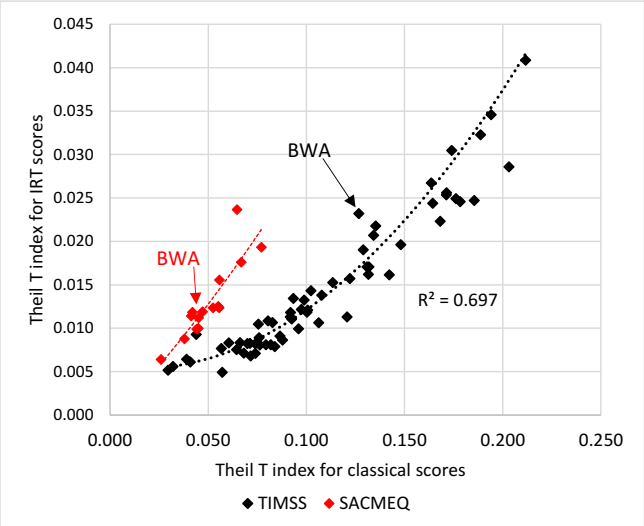
$$I_p = \alpha + \beta C_p \qquad (1)$$

[20] Theil T is the generalised entropy index with parameter 1.0

Here C is the classical score. Parameter values of 260 for α and 614 for β is what one obtains in a regression using Botswana's 2007 SACMEQ data. IRT scores calculated in this manner would produce a Theil T value of 0.0117, compared to a Theil T of 0.0471 for the underlying classical scores C. The fact that α makes such a difference illustrates what is known as the translation invariance of all our four inequality measures. The difference between 0.0117 and the 0.0119 value for SACMEQ Botswana on the vertical axis is mostly attributable to the non-linearity of the classical-IRT relationship seen in Figure 4.

Given the incomparability, the same inequality measure based on different test scores, it should not come as a surprise that comparisons with country-specific measures of income inequality are not meaningful. For example, the Gini coefficient for Botswana's 2007 SACMEQ scores is 0.17 if classical scores are used, and 0.09 if IRT scores are used. The World Bank estimates Botswana's Gini coefficient for income to be around 0.60. Clearly one cannot claim that income inequality is four or six times as high as inequality with respect to test scores. In fact, it is not inconceivable for a differently designed test to produce an education Gini of 0.60. Such a test would have to allow a few pupils to score full marks, but half of the pupils would have to score zero, and a further quarter would have to score no more than one correct out of 100. The TIMSS classical scores for Botswana actually produce a Gini coefficient of 0.29, or half of the income Gini. Any measure of test score inequality is highly sensitive to the structure of the test, and the score type. In that sense, one might be able to argue that neither classical nor IRT scores are as strongly and "simply" cardinal as income.

Figure 11: Classical versus IRT scores and measures of inequality



Note: Just one outlying marker is not shown, that of Yemen, whose x and y values are 0.182 and 0.076. The most recent year's data per country for each of the testing programmes were used.

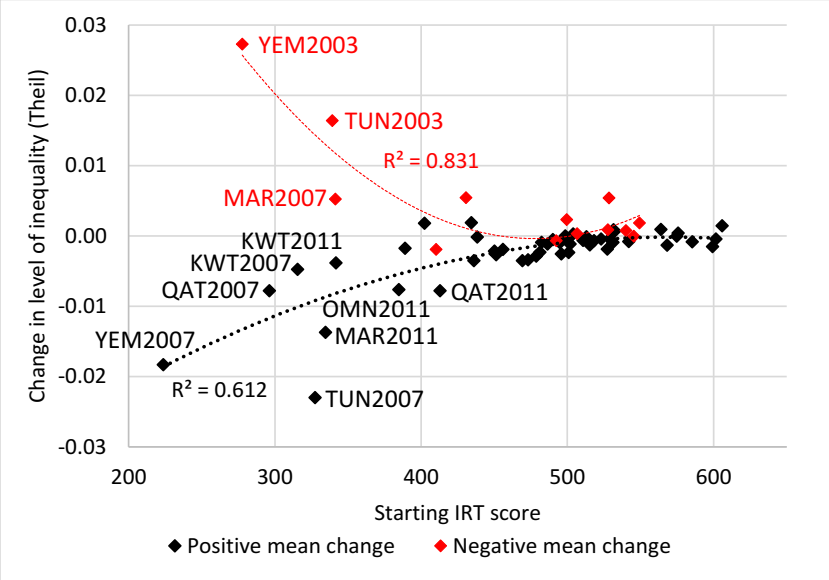
Table 6 below examines the ninety-three instances of mean score change. We see that in around a third of these instances inequality worsened if we use IRT scores, and that this becomes just under half if we use classical scores (third and sixth columns). Either way, most of the movement was towards less inequality. Moreover, whichever score type one uses, the general pattern was for mean 'decliners' (which could be with respect to the IRT or classical mean) to be most prone to a worsening in inequality. For example, of forty instances of a decline in the classical mean, thirty-two involved an increase, or a worsening, in the Gini coefficient, whilst of fifty-three improvers, just twelve saw their Gini coefficient worsen.

Table 6: More or less inequality?

	IRT			Classical		
	Improver	Decliner	All	Improver	Decliner	All
Total	70	23	93	53	40	93
Number within the total where inequality *worsened*						
Theil T	14	17	31	11	30	41
GE -1.0	16	18	34	16	13	29
Gini	13	5	18	12	32	44
p90/p10	16	16	32	16	28	44

Figure 12, a critical graph in the analysis, reflects the sixty-one two-point time series discussed previously (these sixty-one are a sub-set of the ninety-three). Thus, each point represents an instance of significant improvement or deterioration in a country's TIMSS mean from one year to another. The analysis points to a clear and negative correlation between an improver country's level of TIMSS performance and change in the level of inequality, as measured by the Theil T index (other measures of inequality are explored below). Put differently, equality gains are the largest for improvers with a low mean at the outset. To illustrate, of the eight instances of TIMSS improvements seen in Figure 9 where the starting point was less than 400 points, all eight were associated with reductions in inequality. The three instances where the TIMSS score declined, and the point of departure was below 400, were associated with worsening inequality.

Figure 12: Changes in measures of inequality (TIMSS I)



Note: Years refer to the starting year in the time series, which always spans four years. No point qualifying for inclusion in this graph had a timespan exceeding the single TIMSS cycle of four years. All changes reflected on the vertical axis are thus changes over four years. The trendlines are quadratic.

To gain a sense of the magnitude of the equality gains illustrated by Figure 12, one can consider a weakly performing, but improving, country such as Yemen in 2007. With effective education policies, and of course an absence of war, one could expect such a country to reduce its level of inequality by a margin equal to one-eighth of the inequality gap between the most unequal and most equal countries of the world. This should be achievable over four years. For the global inequality gap, the difference between the 10^{th} and 90^{th} percentiles amongst the TIMSS countries in Figure 11 was used. This link between better average performance and greater equality is what Freeman *et al* (2011) refer to as a virtuous efficiency-equity link.

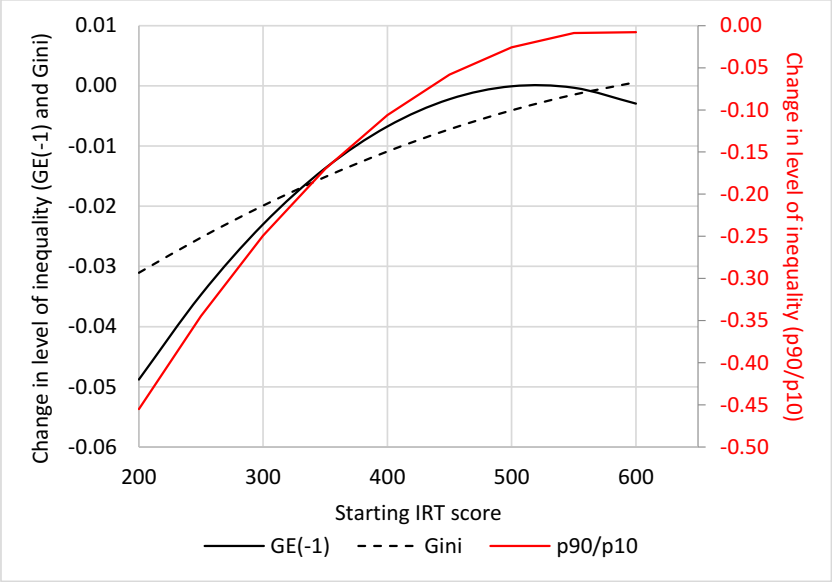
The following multivariate regression was run to explore the three-way relationship between initial TIMSS level of performance, the size of the positive or negative TIMSS score change and change in the level of inequality. Here the initial TIMSS score (T_1) for country *i* and the change in the TIMSS score are regressed on the change in the value of the Theil T index (*t*). The initial score *T*i is entered twice and in each instance multiplied by a dummy variable *D* indicating whether the change in the TIMSS score was positive (D_a) or negative (D_b). The number of observations was ninety-three, representing the ninety-three two-period time series referred to previously. This reduces somewhat a problem with Figure 12, namely that trends were based on few observations. In the multivariate regression the coefficient β_3 is

significant at the 1% level, whilst β_1 and β_2 are not significant at the 10 percent level. Thus despite the apparent sensitivity of the change in inequality to the initial level of TIMSS performance seen in Figure 12, in this more comprehensive analysis it is the size of the performance change, not initial performance, which correlates with a change in inequality (as measured by Theil T). Striving to improve the average test score appears highly complementary to reducing educational inequality.

$$\left(t_{i2} - t_{i1}\right) = \alpha + \beta_1 D_a T_{i1} + \beta_2 D_b T_{i1} + \beta_3 \left(T_{i2} - T_{i1}\right) \qquad (2)$$

Figure 13 below reproduces the black curve from the previous graph, using alternative measures of inequality, specifically the generalised entropy index with a parameter of -1.0 (this parameter makes the measure particularly sensitive to changes at the lower end of the distribution), the Gini coefficient and performance at the 90th percentile divided by performance at the 10th percentile. The three alternative measures essentially reproduce the conclusions found using the Theil T index. In particular for less developed countries, the predominant trend is for improvements in the TIMSS mean to be associated with a reduction in inequality.

Figure 13: Changes in measures of inequality (TIMSS II)

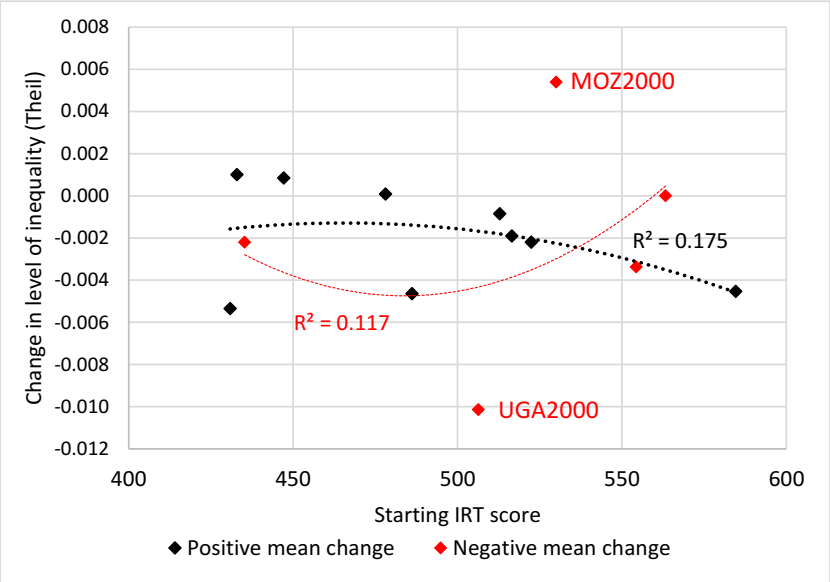


Note: The red curve should be read against the right-hand vertical axis. The adjusted R^2 value for the three curves are 0.692, 0.418 and 0.584.

As seen in Table 4, measures of within-country inequality are less stable over time in the SACMEQ data compared to the TIMSS data. This is probably because SACMEQ gauges pupil performance less consistently than TIMSS. One might not expect the patterns of Figure

12 to be repeated if one uses the SACMEQ data. As seen in Figure 14 below, they are indeed not replicable. Yet for nine of fourteen SACMEQ countries the move was towards less inequality.

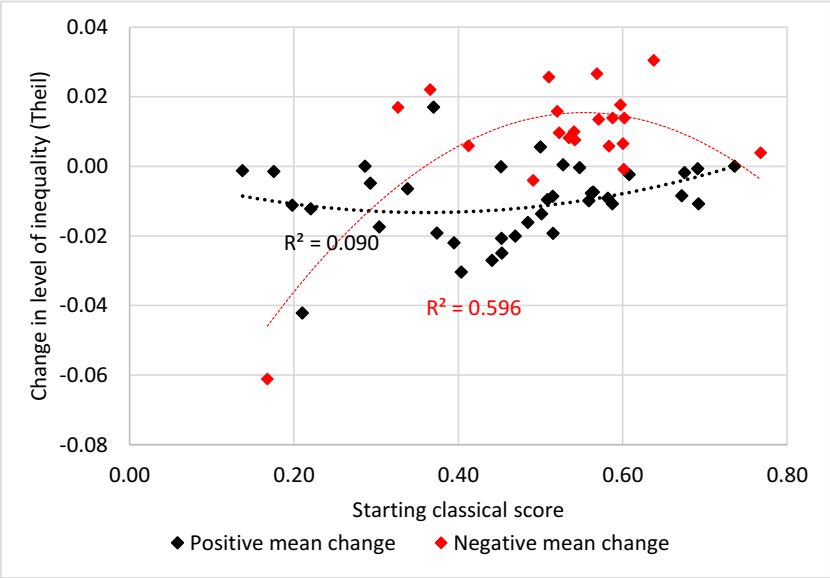
Figure 14: Changes in measures of inequality (SACMEQ)



Note: Change in the level of inequality is over seven years (2000 to 2007).

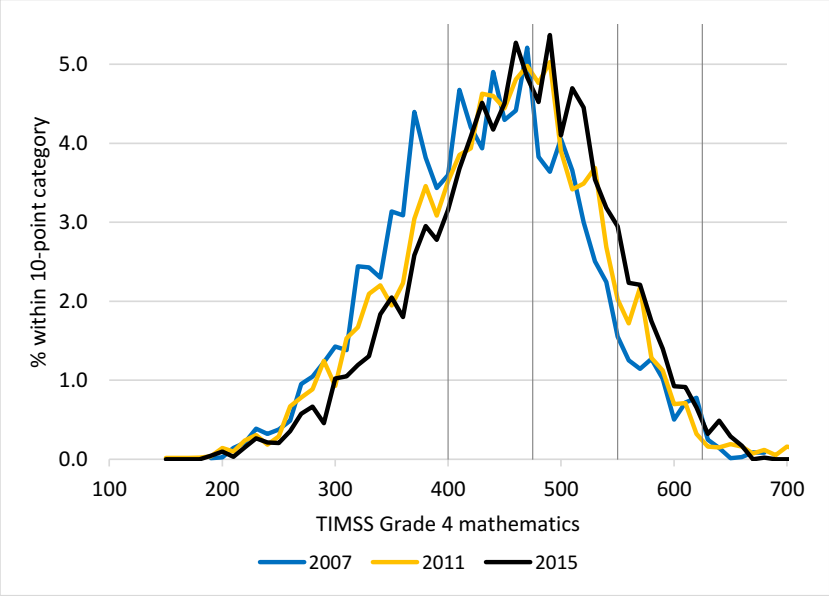
Figure 15 repeats Figure 12 using TIMSS classical scores. Again, patterns are not replicated. There is no clear link between the level of development of improvers and their change in inequality. There is also no support for a Kuznets effect: improvers towards the left of the graph are not more likely to experience an increase in inequality. Comparing Figure 12 to Figure 15 moreover suggests that IRT scores are better at producing meaningful measures of inequality, and especially measures of the change in inequality, than classical scores.

Figure 15: Changes in measures of inequality (TIMSS classical)



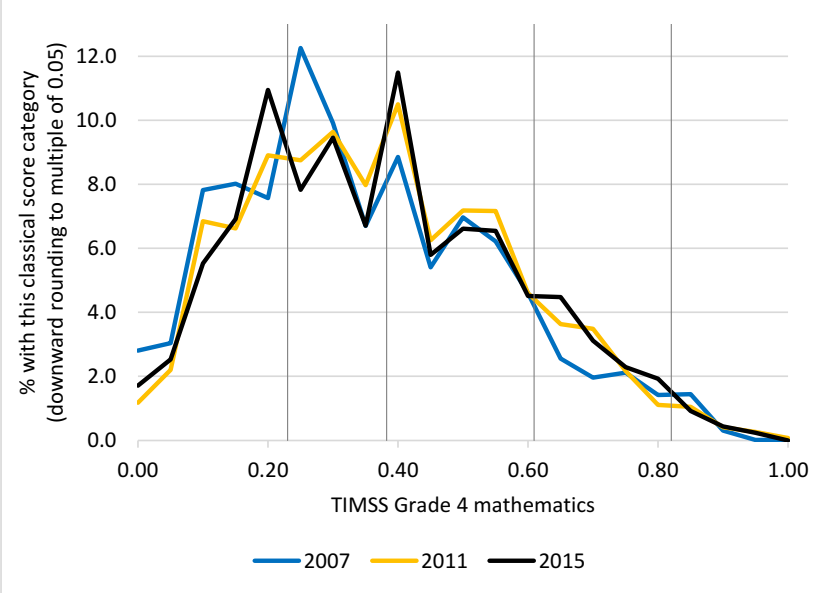
The following two graphs and one table explore the difference between the IRT-based analysis and the classical-based analysis in more depth, with reference to Georgia, a country with a 2015 TIMSS mean of 463, which improved, and which saw reductions in inequality. The problem of floor effects in the classical score data is clear from Figure 17. The percentage of pupils with a classical score of zero (before any rounding used to produce the graphs), was 0.6 percent, 0.2 percent and 0.5 percent for the three years covered. The IRT-based inequality trend seen in Table 7 is continuous, against a down and then up trend when classical scores are used. It is unlikely that inequality in educational performance would in fact be jumpy. All this seems to confirm the superiority of IRT scores for gauging within-country inequality. Moreover, one reason for this appears to be the superior ability of the IRT scores to differentiate amongst pupils at or near the classical floor.

Figure 16: IRT score distribution in Georgia over time



Note: Vertical lines represent the cut-offs for the TIMSS international benchmarks.

Figure 17: Classical score distribution in Georgia over time



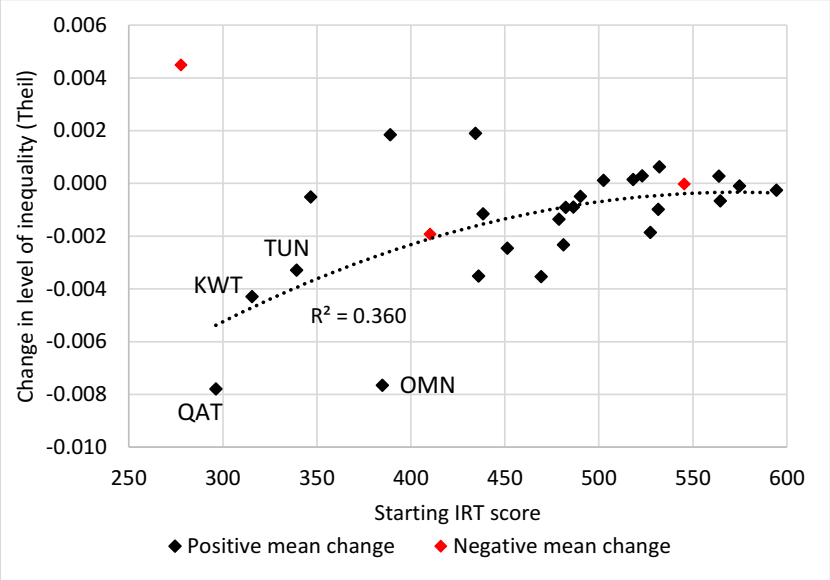
Note: Vertical lines represent classical score equivalents of the international benchmarks.

Table 7: Trends in Georgia

	2007	2011	2015
Using IRT scores			
Mean	438	450	463
Gini	0.109	0.107	0.100
Theil T	0.019	0.018	0.016
Using classical scores			
Mean	0.37	0.40	0.40
Gini	0.299	0.279	0.286
Theil T	0.145	0.125	0.132
Note: IRT mean values are obtained using weighted pupil values. They are the same as the official means published by TIMSS.			

Figure 18 provides further evidence of the robustness of the patterns seen in Figure 12. Here each point represents just one country. For instance, Qatar's trend from 2007 to 2011 to 2015 is represented by one and not two points (as was the case in Figure 12). The change in the level of inequality is the annual slope for the Theil T measure across the two or more years. The trendline in Figure 18 is similar to that of Figure 12 (values on the vertical axis of Figure 18 must be multiplied by four to make them comparable to those of Figure 12).

Figure 18: Changes in measures of inequality (TIMSS III)



Note: 26 countries with a positive mean change exceeding 1.5 TIMSS points per year are included, and three countries with a negative mean change per year exceeding -1.5. Annual change is the slope across two or more points in time.

4. Inequality decomposition

To end the discussion of inequality, the impact of score type on inequality decompositions is briefly examined, using both micro data and an approximate method using macro data.

Using the micro data, whether one uses IRT or classical scores makes a large difference to conclusions around how much of the educational inequality exists across countries, as opposed to within countries. In producing Table 8, pupil weights in the data were converted to percentages of all pupils within each country, meaning each country was given an equal weight. Moreover, for each country, only the most recent TIMSS data were used. The within-country and between-country Theil T values in Table 8 add up to the overall Theil T measure of inequality. For TIMSS IRT-based measures, the bottom line is that 46.4 percent of the overall inequality found in TIMSS is between countries. This is close to the 49 percent found by Sahn and Younger (2007: 11) using TIMSS Grade 8 IRT scores, and the Theil T (or GE[1]) measure. However, switching to TIMSS classical scores produces a very different breakdown, of 32.3 percent. SACMEQ, on the other hand, produces a very similar between-country percentage across the two score types, though this is likely to be in part due to the fact that SACMEQ IRT scores are not truly IRT scores (due to the one-to-one problem discussed above). The conclusion that can be drawn is that just as one cannot extract much

information from any single education inequality value, one should not read too much into single decompositions of test score inequality.

The last row of Table 8 reflects the percentage of variance within countries, using the approach employed, for instance, by the OECD to distinguish between within-school and between-school test score variances in PISA.[21] The results are very close to those obtained using the Theil T measure.

Table 8: Inequality decompositions

	TIMSS (64 countries)		SACMEQ (15 countries)	
	IRT	Classical	IRT	Classical
Decomposition using Theil T				
Overall	0.0266	0.1420	0.0175	0.0673
Within-country	0.0142	0.0961	0.0132	0.0505
Between-country	0.0123	0.0459	0.0043	0.0168
% between	46.4	32.3	24.7	25.0
Decomposition using variance				
% between	46.2	33.2	24.9	25.0

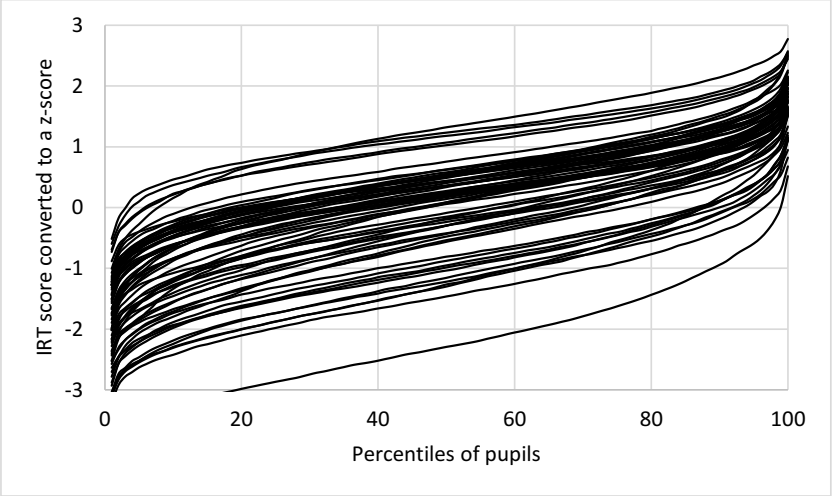
If the TIMSS values from Table 8 are reproduced using the other inequality measure (of our four measures) which allows for a decomposition by level, namely the generalised entropy index with parameter -1.0, the finding remains essentially the same. The percentage of inequality which exists between countries is 41.3 percent if IRT scores are used, and 20.0 percent if classical scores are used. Values change by a large margin depending on what scores one uses.

The following two graphs illustrate why IRT scores produce more between-country inequality than classical scores in Table 8 above. For these graphs, which focus only on TIMSS, scores were first converted to z-scores with a mean of zero and a standard deviation of one. IRT scores differentiate to a larger degree across countries with respect to the very weakest performing pupils and with respect to the best performing pupils. To illustrate, at the very bottom end, a pupil obtaining a classical score of zero is likely to be given a better IRT score in, say, Singapore compared to such a pupil in Yemen. This is because, for instance, the

[21] See for instance OECD (2004: 162).

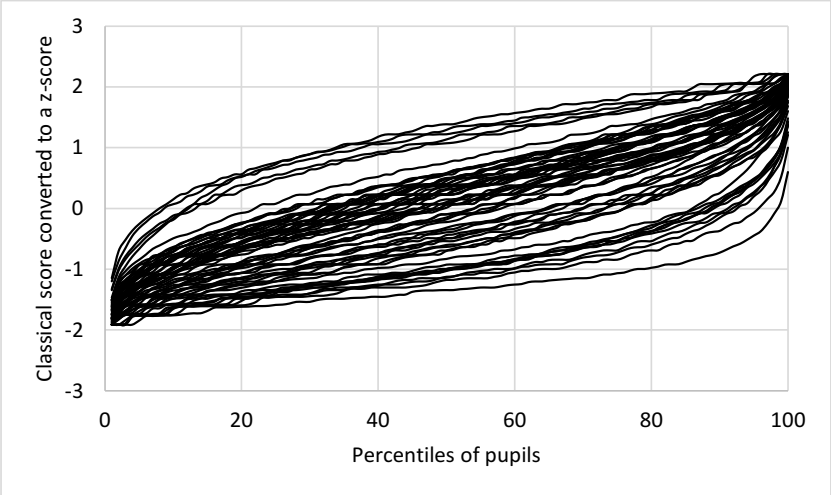
pupil in Singapore is less likely to have scored zero on *both* the mathematics and science tests. This largely explains the higher between-country inequality when IRT scores are used.

Figure 19: Country distributions of TIMSS IRT scores converted to z-scores



Note: Each curve in the current graph and the next graph represents a country.

Figure 20: Country distributions of TIMSS classical scores converted to z-scores



Turning to macro data, it is possible to produce a simple and intuitive measure of inequality, and its decomposition, based on the notion of differences between scores at the 95th and 5th percentiles of a distribution. The measure is not inherently decomposable into within- and between- components, but it turns out, empirically, to be approximately decomposable.

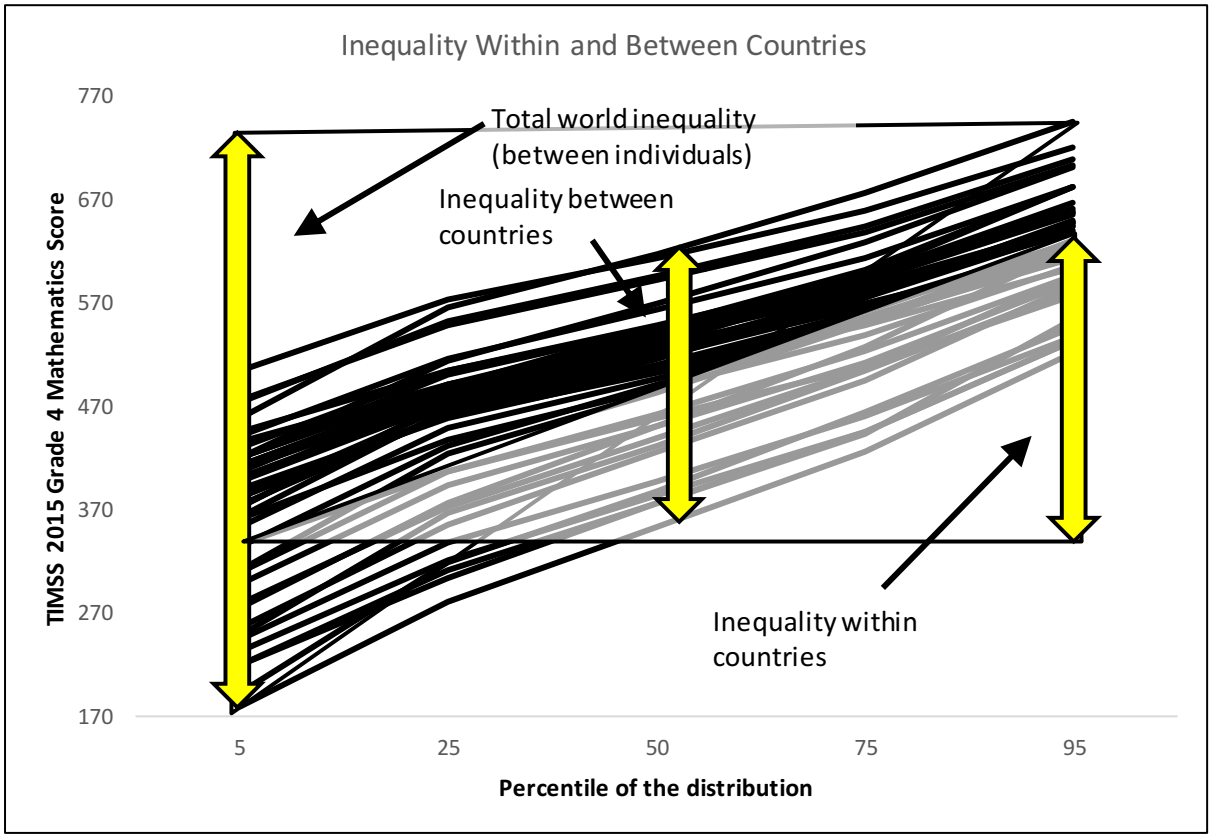
The inspiration for the measure can be seen in Figure 21. Here, total global inequality approximated as follows. Take a), the score of the student at the 95th percentile of the

country at the 95th percentile of the between-country distribution, namely the "top" student in the "top" country. Now take b), the score of the student at the 5th percentile country at the 5th percentile of the between-country distribution, namely the "bottom" student in the "bottom" country. The difference a) minus b) is global inequality, and is signified in the graphic by the length of the longest, left-most arrow. Typical within-country inequality is the difference between the student at the 95th percentile in the median country (the country with the middle line in the set of lines) and the student at the 5th percentile in the median country. This is represented by the right-most arrow. Finally, between-country inequality is the difference between the median student in the country at the 5th percentile of the countries (the score of the student at the 50th percentile, on the horizontal axis, in the lowest line), and the score of the median student (the one at the 50th percentile in the horizontal axis) in the country at the 95th percentile of the countries (the highest line). This is represented by the middle arrow. Note that if we were interested in the absolute value of these measures, it would make sense to normalize by the overall median, to create a sort of coefficient of variation, something a little less unit-variant. But given that our purpose here is decomposition, a denominator would cancel out when calculating the shares.

To calculate our measures, we take the share of "between country inequality" to "total inequality" to be the ratio of the length of the middle arrow to the sum of the middle and right-most arrows. The share of "within country inequality" to "total inequality" is calculated as the ratio of the length of the right-most arrow to the sum of the middle and right-most arrows. To calculate whether these shares are approximately decomposable, the ratio of the sum of the middle and right-most arrows to the left-most arrow is also calculated. This is presented in Table 9 as a "decomposability index."

An intuition as to why the measure of inequality is not decomposable by construct is that if the parallelogram implicit in the figure below had equal left-hand lower and right-hand upper angles, then the measure would be decomposable exactly and by construct. But these parallelograms do not have that property, not exactly, though they come close.

Figure 21. TIMSS 2015 Grade 4 mathematics approximation to between- and within-country inequality



Calculating these data for all the assessments we are considering, using macro, country-level data, produces Table 9.

Table 9. Between-country inequality and deviation from decomposability					
Assessment	Between-country inequality	Decomposability	Assessment	Between-country inequality	Decomposability
PISA 2015 Reading	0.36	1.09	PIRLS 2011	0.36	0.94
PISA 2015 Mathematics	0.37	1.02	SACMEQ III Reading	0.34	1.13
PISA 2015 Science	0.39	1.00	SACMEQ III Mathematics	0.33	0.93
TIMSS 2015 Grade 4 Mathematics	0.48	0.95	TERCE Grade 3 Reading	0.33	1.00
TIMSS 2015 Grade 8 Mathematics	0.44	0.96	TERCE Grade 3 Mathematics	0.34	0.95
TIMSS 2015 Grade 4 Science	0.43	0.86	TERCE Grade 6 Reading	0.30	1.06
TIMSS 2015 Grade 8 Science	0.38	0.97	TERCE Grade 6 Mathematics	0.38	0.96

The shares are simply the proportion of total inequality that is accounted for by the between-component, measured as explained above. The “decomposability” is simply the ratio of the sum of the between- and within- components to the total inequality. As can be noted from the table, for most of the assessments the distance between the sum of the components and the total inequality is not far, so the idea of decomposing in this approximate manner would seem reasonably acceptable as a first approximation.

Note, comparing TIMSS 2015 Grade 4 mathematics and SACMEQ mathematics, that the between- components as calculated using the micro data in Table 8 and Table 9 are fairly similar: 46 percent and 25 percent for the two assessments using micro data (IRT version), and 48 percent and 33 percent using the macro data.

It is logical that the regional assessments would show less between- inequality: they naturally have fewer country-mean extremes.

We do not know why there is higher between- inequality shown in TIMSS, other than a somewhat circular explanation: TIMSS (taking Grade 4 mathematics) shows much higher performance at the upper end than does, say PISA Science: 600 versus 531 at the 95th percentile of either assessments country-mean distribution.

Taking the international assessments only, because they likely represent something closer to worldwide variation (inequality), the "between-" component averages to about 40 percent (and hence the "within-" component would be about 60 percent).

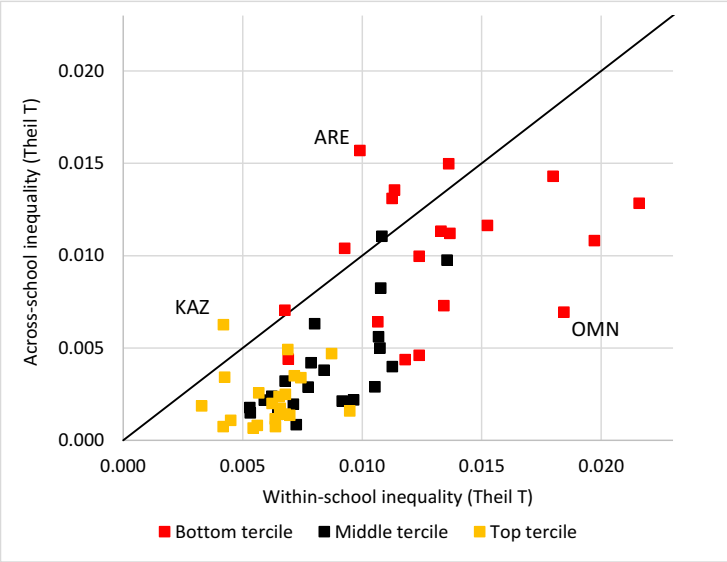
It is likely that if more developing countries were represented in the PISA, TIMSS, and PIRLS samples, and especially if we used population-weighted averages, the "between-" component would grow. Other analysts have tried to develop a sense of how much things would change by creating predicted values for countries that do not participate in the samples. We tried something similar, but the prediction basis, using factors such as GDP per capita, was so weak, that we gave up. Sorting this out would be an exercise for a paper that did nothing but that.

5. Schools that reach at least the foot of the pyramid

Whilst the principal units of analysis in the paper are the pupil and the country, applying the data and methods used so far to an examination of between-school inequalities within countries helps to shed light on the policy solutions needed to improve learning.

Figure 22 below displays the results of within-country decompositions, using the Theil T measure of inequality, across schools. More developed countries, in terms of their TIMSS performance, clearly display less within-school and between-school inequality in absolute terms. Only nine countries, of which seven are in the bottom performance tercile, display more between-school inequality than within-school inequality. This serves as a reminder of the importance of certain policy measures in, above all, developing countries. These countries need systems to identify which schools perform below expectations, and systems to promote effective and accountable school managers.

Figure 22: Within- and between-school inequality using TIMSS IRT scores



Note: The most recent year's data for each country were used. Pupil weights were used. The 64 countries were divided into terciles, based on their average TIMSS scores.

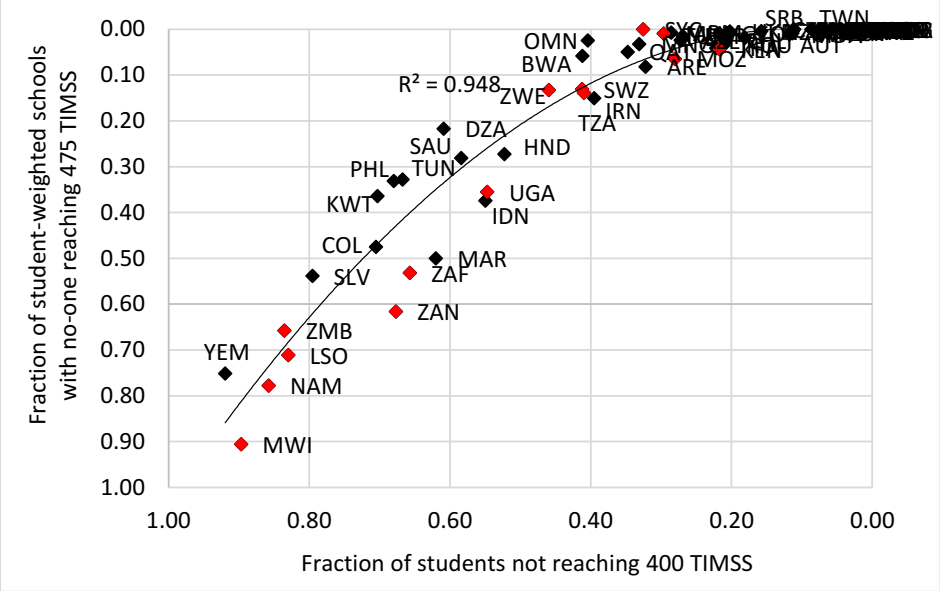
How does picture provided by the above graph change if classical scores are used, instead of IRT scores? In line with the patterns seen earlier when within- and between-country inequalities were considered, the use of classical scores reduces the proportion of total within-country inequality accounted for by between-school inequality. Once again, the use of classical scores suppresses measures of 'between' inequality. In fact, using classical scores would produce a distribution without any schools above the diagonal in the graph. No countries would display more between-school inequality than within-school inequality. Yet the pattern of between-school inequality being a larger problem in developing countries remains. This reinforces the overall finding of the paper that analyses of educational inequality are highly sensitive to whether the original classical scores are used, or IRT scores.

The following graph presents an approach to understanding between-school inequalities which might be more practical to policymakers than the previous graph. Minimum thresholds in relation to cognitive skills at specific levels of education are useful insofar as they can guide policymakers in relation to the attainment of education as a basic human right. This is in part the approach of the Sustainable Development Goals, where reference is made to indicators such as 'Percentage of girls and boys who master a broad range of foundational skills, including in literacy and mathematics, by the end of the primary school cycle'.[22]

[22] United Nations: Sustainable Development Solutions Network, 2015: 139.

Hanushek and Woessman (2007: 56) refer to the non-attainment of 'functional literacy' as an impediment to economic growth. In Figure 23 below, the horizontal axis refers to a typical threshold, in this case the percentage of Grade 4 children not reaching at least the low international benchmark of 400 TIMSS points. Here SACMEQ values were recalibrated to the TIMSS scale following the 'equipercentile linking' approach used by Sandefur (2016) to link SACMEQ to TIMSS Grade 8 data. In the linking implemented for Figure 23, Botswana's 2011 Grade 6 TIMSS data[23] were linked to Botswana's 2007 SACMEQ Grade 6 data, in both cases focussing just on mathematics.

Figure 23: The upward pathway and school functionality



Source: TIMSS and SACMEQ mathematics microdata.

Note: The most recent year's data for each country were used. Red markers refer to SACMEQ values recalibrated to TIMSS values. The trendline is quadratic.

The vertical axis presents a novel indicator, focussing on across-school inequalities, which seems worth considering. The operations of any education authority centre around schools, not pupils. The more these authorities are able to characterise schools using the data at their disposal, the greater the practical uses of the data. While partitioning inequality within a country into between-school and within-school inequality is fairly routinely done,[24] using international test data to generate meaningful school-level performance thresholds, along the lines of the pupil-level thresholds described above, seems unexplored territory. The measure along the vertical axis of Figure 23 is the percentage of children in schools where

[23] Botswana was one of three countries in 2011 to test grade 6 instead of grade 4 (the other two were Yemen and Honduras).
[24] See for instance OECD (2004: 162).

not even *one* pupil reaches a threshold one might consider a bit ambitious, but certainly not impossible. A threshold of 475 was used, this being the minimum for the TIMSS intermediate level, and slightly higher than the median score of 469 for the country of Georgia in 2015.

The patterns seen in Figure 23 lend themselves to intuitive and practical policy conclusions. One noteworthy pattern is that by the time countries reach an educational 'poverty level' of 20 percent, meaning that 20 percent of pupils score less than the low international threshold, they are left with virtually no schools where no pupil achieves the higher threshold of 475. Such schools, it could be argued, display an easily measurable dysfunctionality. Even with serious home background obstacles impeding pupils, it seems unlikely that a relatively well-managed school should have *no* children reaching a level such as 475.

To illustrate the utility of the graph, one can compare the neighbouring countries of Algeria (DZA) and Morocco (MAR). Both have similar 'poverty incidences' in the sense that in both countries around 60 percent of pupils do not reach the low international threshold. Yet Algeria seems far better positioned to tackle this poverty than Morocco insofar as its school-level inequalities are smaller. Specifically, whilst in Algeria around 20 percent of pupils are in schools with such a low level of functionality that no pupil reaches a level of 475 points, in Morocco around 50 percent of pupils find themselves in such dysfunctional schools. Morocco needs to pay even more attention than Algeria to ensuring that schools reach a minimum level of functionality.

The graph serves as a reminder that gauging the functionality of schools is not just a question of focussing on school-level averages, which is the typical approach, but also on the performance of each school's best pupils. This might appear to run counter to a pro-equity strategy. A policy aimed at ensuring that every school produces at least one pupil with a relatively high test score could well increase within-school inequality. However, such a policy would support equity insofar as it widened the conduit from less advantaged segments of the system to post-school studies leading to better-paying occupations. Put differently, it would contribute towards a less rigid cross-generational transfer of human capital.

The magnitudes revealed by Figure 23 are telling. To move from a situation where almost 100 percent of pupils score below 400 TIMSS points to a point where about 40 percent score below this level, implies a large reduction in the proportion of dysfunctional schools. Specifically, the percentage of such schools needs to be reduced from around 100 to 10 percent .

In economics, there is a tradition of viewing a country's readiness for economic growth in terms of its 'frontier technologies', or its most advanced industries.[25] These frontiers offer opportunities for within-country diffusion of more efficient production processes. Analogously, one can think of the presence of one or two exceptionally good teachers in an otherwise weak school, or the presence of a dynamic school principal in such a school, as 'frontier technologies' serving as catalysts for change, and raising the probability that one or two pupils will excel.

6. Conclusions, possible policy implications, ideas for further research

The paper uses both macro (published, country-level) data from the most recent wave of all the major international assessments, for all grades and all subjects, and micro data from TIMSS 2015 Grade 4 mathematics and SACMEQ III mathematics, to explore a set of issues regarding inequality and cognitive poverty around the world.

Our conclusions are that:

1. As against other research that uses one or a few assessments only, we do not see evidence that inequality, as measured by the standard deviation of results across countries, is negatively or positively correlated with mean country achievement. Thus, we do not see evidence that inequality either increases or decreases as means increase. In some assessments it increases, in others it decreases. It is a bit of a cause for worry that this pattern is dependent on which institution produces the assessments, and this is an area for further analysis.
2. While there is little evidence about the direction of inequality in response to increases in mean performance, there is very strong evidence about what happens to cognitive poverty, or the percentage of students at very low levels of achievement, as mean country performance improves: it decreases strongly between very low country mean performance and median mean country performance, then it decreases a bit more, but much more weakly, between median mean country performance and high mean country performance. The percentage of students at high levels of performance is exactly symmetrically opposite: countries improve the percentage of students at the higher benchmarks only after they are past the median of mean country performance. This assumes that cross-section can be proxy for time dynamics.
3. The evidence using time dynamics, either produced for this paper, or produced by others, but referenced here, however, suggests the same thing: the percentage of

[25] Nelson and Phelps, 1966.

students at very low levels of performance decreases with increases in mean country performance over time.

4. International assessments allow (and national ones would allow) one to estimate the percentage of schools where not even a single student rises above a relatively low "cognitive poverty" threshold (here defined as 475 in TIMSS 2015 Grade 4 mathematics, and translated to its equivalent in SACMEQ III). There are big differences between countries on this score, and it is suggested that countries where the number of schools where at least one student rises above a certain level might have an easier time improving their mean scores. To the degree that this is a type of inequality, this type of inequality might be useful.
5. Finally, an exercise in decomposing cognitive inequality into between-country and within-country components is carried out, using both micro (student-level) and macro, published, country level data, with the full set of assessments. The results are more or less consistent with each other and with literature that uses only one or only a few assessments. About 60 percent of inequality is within countries, and "only" 40 percent is between countries. The "between-" component would likely grow if the international assessments sampled more (large) developing countries.

Policy implications are fairly clear. Poorer countries, namely those of interest, typically, to the RISE Programme, DFID, and international development agencies in general, can help themselves most, if measures are taken to cut down the percentages of students performing at very low levels. This seems to be the way that countries progress from very low levels of overall performance to at least middling levels of overall performance. The pedagogical, systems efficiency, accountability and other implications are beyond the scope of this paper, but literature is emerging on how countries can help themselves to reduce the proportion of students showing extremely low performance.

Furthermore, if it is true that so much of inequality is within countries, and assuming a goal of reducing worldwide inequality, it makes sense, from an international development agency point of view, to help countries deal with the "within-country" component of inequality as opposed to overall improvement of poor countries in relation to well-off countries.

All this is not to minimize the importance of improving mean levels of cognitive achievement. This is also important, for many reasons, including its connection to economic growth. It is to say, however, that in fact the improvements in the mean, from a low base, seem to require that deep cognitive disadvantage be addressed.

Areas for further analysis suggest themselves. One could carry out micro research with all of the data sets, beyond TIMSS 2015 Grade 4 mathematics and SACMEQ III mathematics, to confirm (or question) the macro results both for the “improve from the bottom” and the inequality decomposition analysis. The fact that the inequality (but not the “cognitive poverty”) results depend so much on differences between particular institutional purveyors of assessments could be looked into further. Perhaps the most interesting area for research, for the RISE Programme, would be to examine precisely what it is that the countries that have progressed by vastly reducing the percentage of children performing very poorly have done. Is it mostly through accountability measures? Through much more adroit pedagogical support to teachers teaching the cognitively disadvantaged, including tighter curricular specification, curricula that do not “shock” children from disadvantaged backgrounds (“teaching at the right level”)? By focusing on the “triple disadvantage” discussed in Crouch and Rolleston (2017)? Carrying out this work would be a major contribution to improving national averages via a reduction in deep cognitive disadvantage.

References

Crouch, L., Gustafsson, M., & Lavado, P. (2009). Measuring educational inequality in South Africa and Perú. In D. B. Holsinger, & W. J. Jacob (Eds.), Inequality in Education: Comparative and International Perspectives (pp. 461-484). Dordrecht, the Netherlands: Springer Netherlands. <DOI: 10.1007/978-90-481-2652-1_20>

Crouch, L. and Rolleston, C. (2017). "Raising the Floor on Learning Levels: Equitable Improvement Starts with the Tail." An Insight note from the RISE Programme. Available at: <http://www.riseprogramme.org/sites/www.riseprogramme.org/files/RISE%20Equity%20Insight_0.pdf> [Accessed on August 2017].

Demery, L. (2000). Benefit incidence: a practitioner's guide. Poverty and Social Development Group, Africa Region, The World Bank. Available at: http://www1.worldbank.org/publicsector/pe/practitioner.doc [Accessed on August 2017].

Dollar, D. and Kraay, A. (2001). "Growth is Good for the Poor." World Bank Working Paper No. 2587. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=632656 [Accessed on November 2017.

Ferreira, F.H.& Giroux, J. (2011.) The Measurement of Educational Inequality: Achievement and Opportunity. IZA Discussion Paper No. 6161. Available at: < http://ftp.iza.org/dp6161.pdf> [Accessed on August 2017]

Freeman, R.B., Machin, S.J.. & Viarengo, M.G. (2011). Inequality of educational outcomes: International evidence from PISA. Regional and Sectoral Economic Studies, 11(3): 5-20.

Bruckauf, Z. & Chzhen, Y. (2016). "Education for All? Measuring inequality of educational outcomes among 15-year-olds across 39 industrialized nations." Office of Research - Innocenti Working Paper WP-2016-08, April. Available at <https://www.unicef-irc.org/publications/pdf/IWP_2016_08.pdf> [Accessed August 2017]

Filmer, D., A. Hasan, and L. Pritchett. (2006). "A Millennium Learning Goal: Measuring Real Progress in Education." Center for Global Development Working Paper 97. <https://tinyurl.com/zbyph5a> [Accessed on August 2017].

Foy, P. & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In Martin, M.O., Mullis, I.V.S. & Hooper, M. (eds.), *Methods and procedures in TIMSS 2015*. Chestnut Hill: IEA. Available from: <https://timssandpirls.bc.edu> [Accessed July 2017].

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

Hanushek, E.A. & Woessmann, L. (2007). *The role of school improvement in economic development*. Washington: National Bureau of Economic Research. Available from: <http://papers.nber.org/papers/w12832.pdf?new_window=1> [Accessed June 2007].

Makuwa, D.K. (2010). Mixed results in achievement. *IIEP Newsletter,* XXVIII(3). Available from: <http://www.iiep.unesco.org> [Accessed October 2010].

Milanovic, B. (2012). "Global Income Inequality by the Numbers : In History and Now." Policy Research Working Paper; No. 6259. World Bank, Washington, DC. Available from https://openknowledge.worldbank.org/handle/10986/12117 [Accessed November 2017)].

Micklewright, J. & Schnepf, S. (2006). Inequality of Learning in Industrialised Countries. IZA Discussion Paper Series, No. 2517. Available from <https://www.researchgate.net/publication/5136840_Inequality_of_Learning_in_Industrialised_Countries> [Accessed August 2017].

Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill: Boston College. Available from: <http://timss.bc.edu> [Accessed December 2016].

Mullis, I.V.S., Martin, M.O., Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum and instruction*. Chestnut Hill: Boston College. Available from: <http://timssandpirls.bc.edu/timss2015/international-results/timss2015/wp-content/uploads/2016/T15-20-years-of-TIMSS.pdf> [Accessed June 2017].

Nelson, R.R. & Phelps, E.S. (1966). Investment in humans, technological diffusion, and economic growth. *The American Economic Review,* 56(2): 69-75.

OECD (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris. Available from: <https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34002216.pdf> [Accessed August 2017].

OECD (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264266490-en.

Oppedisano, V. & Turati, G. (2015). What are the causes of educational inequality and its evolution over time in Europe? Evidence from PISA. *Education Economics,* 23(1): 3-24.

Ross, K.N., Saito, M., Dolata, S. & Ikeda, M. (2008). *Chapter 2: The conduct of the SACMEQ II project*. Paris: IIEP. Available from: <http://microdata.worldbank.org/index.php/catalog/1245/download/22682> [Accessed July 2017].

Sahn, D.E. & Younger, S.D. (2007). *Decomposing world education inequality*. Ithaca: Cornell University. Available from: <http://www.cfnpp.cornell.edu/images/wp187.pdf> [Accessed June 2017].

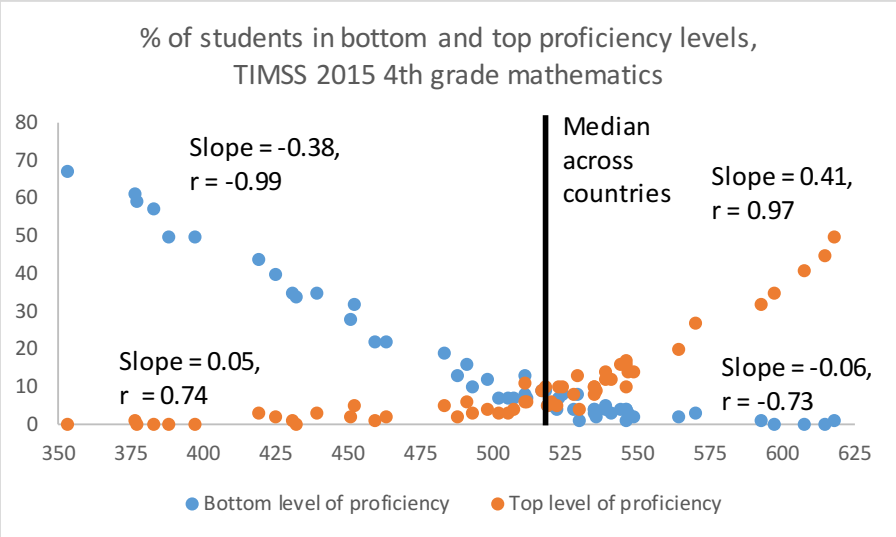
Sandefur, J. (2016). *Internationally comparable mathematics scores for fourteen African countries.* Washington: Center for Global Development. Available from: <http://www.cgdev.org/sites/default/files/math-scores-fourteen-african-countries0.pdf> [Accessed December 2016].

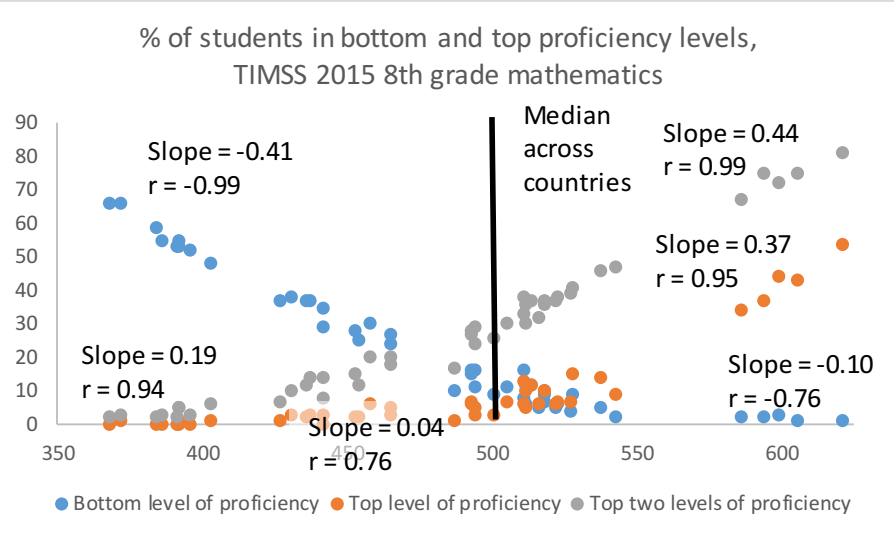
Taylor, S. & Spaull, N. (2015). Measuring access to learning over a period of increased access to schooling: The case of Southern and Eastern Africa since 2000. *International Journal of Educational Development*.

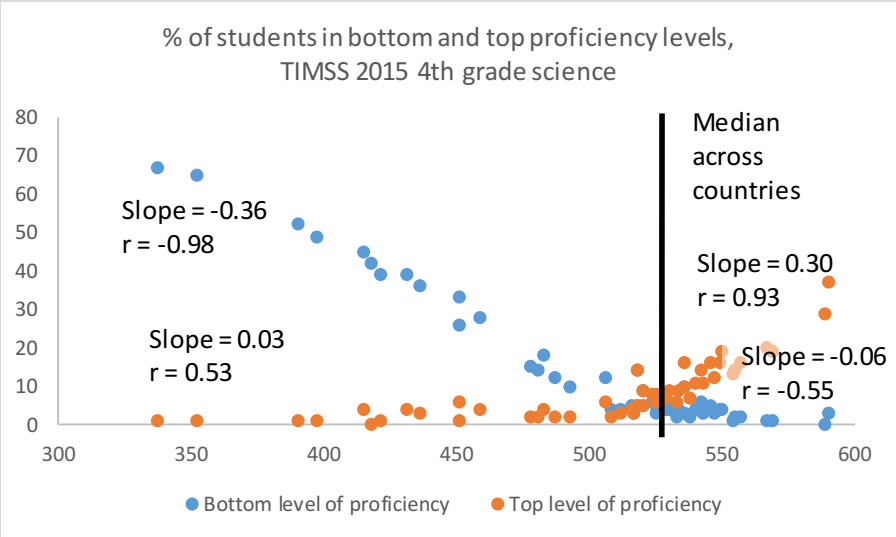
Thomas, V., Wang, Y., and Fan, X. (2001). Measuring educational inequality: Gini coefficients of education, Policy Research Working Paper 2525, Washington: The World Bank Group.

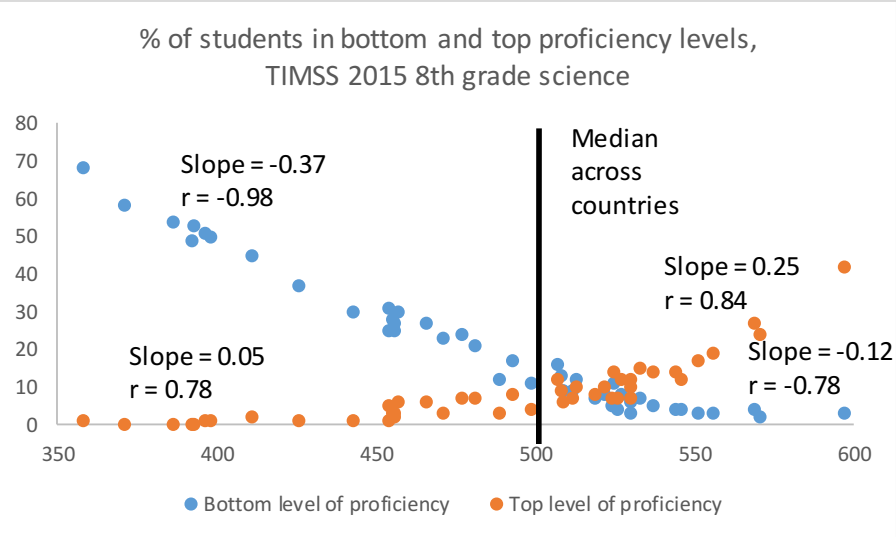
United Nations: Sustainable Development Solutions Network (2015). *Indicators and a monitoring framework for the Sustainable Development Goals.* New York. Available from: <http://unsdsn.org/wp-content/uploads/2015/05/FINAL-SDSN-Indicator-Report-WEB.pdf> [Accessed June 2016].

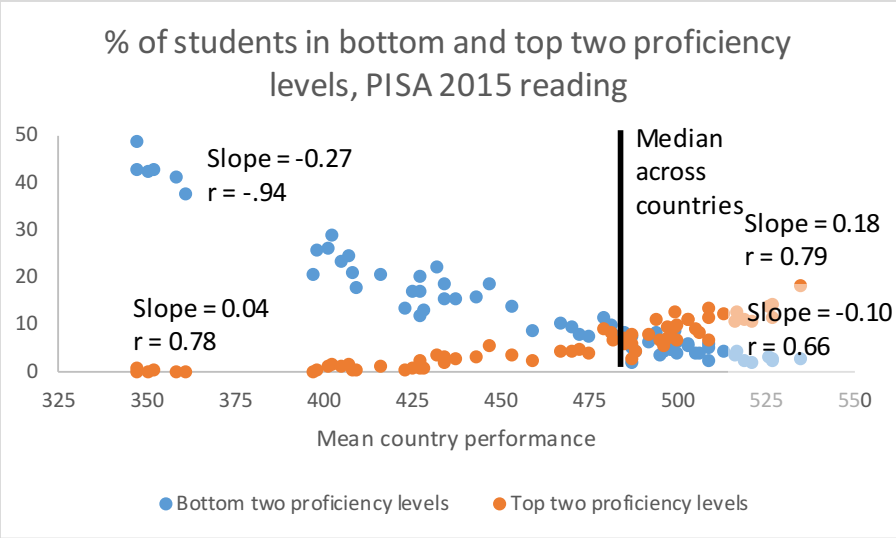
Annex1

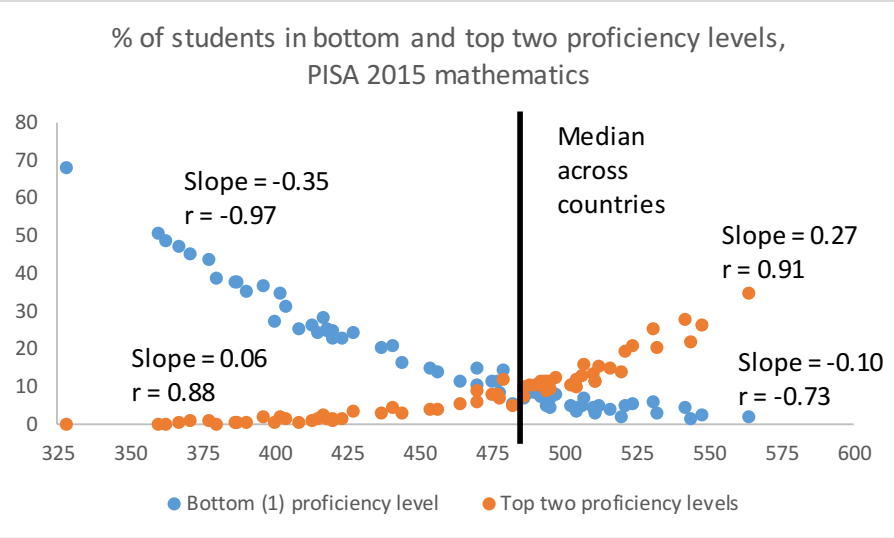

% of students in bottom and top proficiency levels,
TIMSS 2015 4th grade mathematics
Slope = -0.38,
r = -0.99
Median
across
countries
Slope = 0.41,
r = 0.97
Slope = 0.05,
r = 0.74
Slope = -0.06,
r = -0.73
Bottom level of proficiency
Top level of proficiency

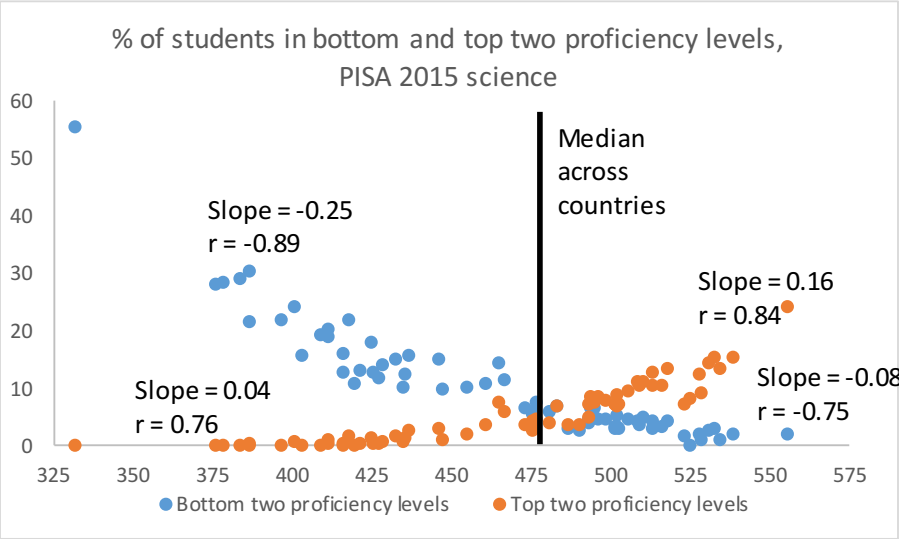

% of students in bottom and top proficiency levels,
TIMSS 2015 8th grade mathematics
Slope = -0.41
r = -0.99
Median
across
countries
Slope = 0.44
r = 0.99
Slope = 0.37
r = 0.95
Slope = 0.19
r = 0.94
Slope = -0.10
r = -0.76
Slope = 0.04
r = 0.76
Bottom level of proficiency
Top level of proficiency
Top two levels of proficiency

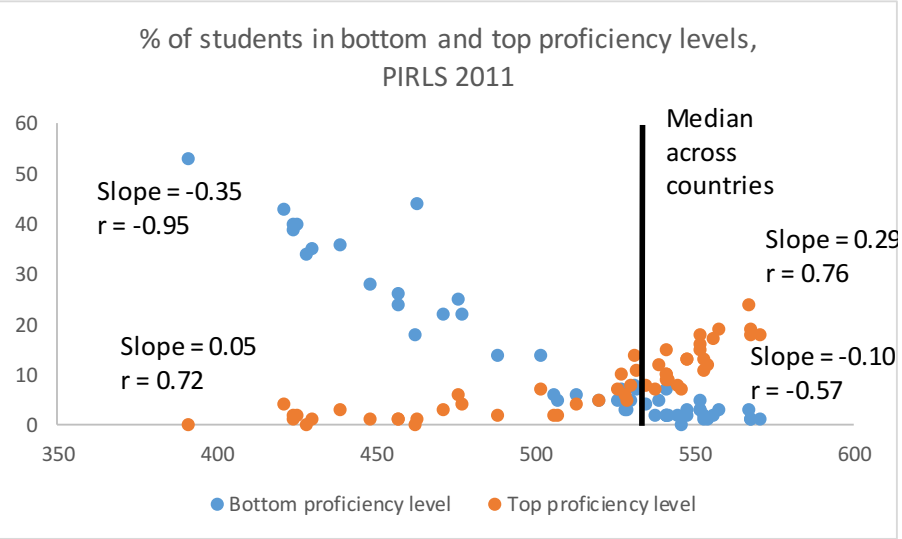

% of students in bottom and top proficiency levels,
TIMSS 2015 4th grade science
Slope = -0.36
r = -0.98
Median
across
countries
Slope = 0.30
r = 0.93
Slope = 0.03
r = 0.53
Slope = -0.06
r = -0.55
Bottom level of proficiency
Top level of proficiency

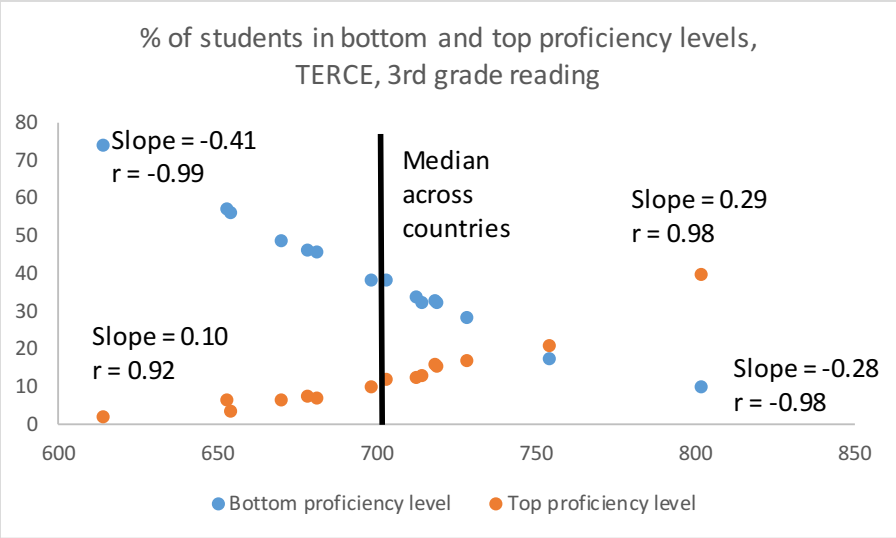

% of students in bottom and top proficiency levels,
TIMSS 2015 8th grade science
Slope = -0.37
r = -0.98
Median
across
countries
Slope = 0.25
r = 0.84
Slope = 0.05
r = 0.78
Slope = -0.12
r = -0.78
Bottom level of proficiency
Top level of proficiency

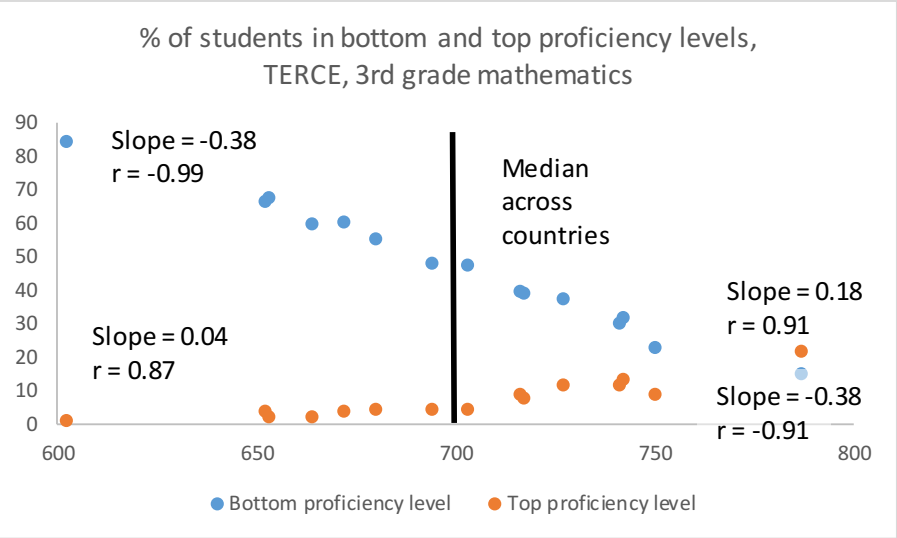

% of students in bottom and top two proficiency levels, PISA 2015 reading
Slope = -0.27
r = -.94
Median across countries
Slope = 0.18
r = 0.79
Slope = 0.04
r = 0.78
Slope = -0.10
r = 0.66
Mean country performance
Bottom two proficiency levels
Top two proficiency levels

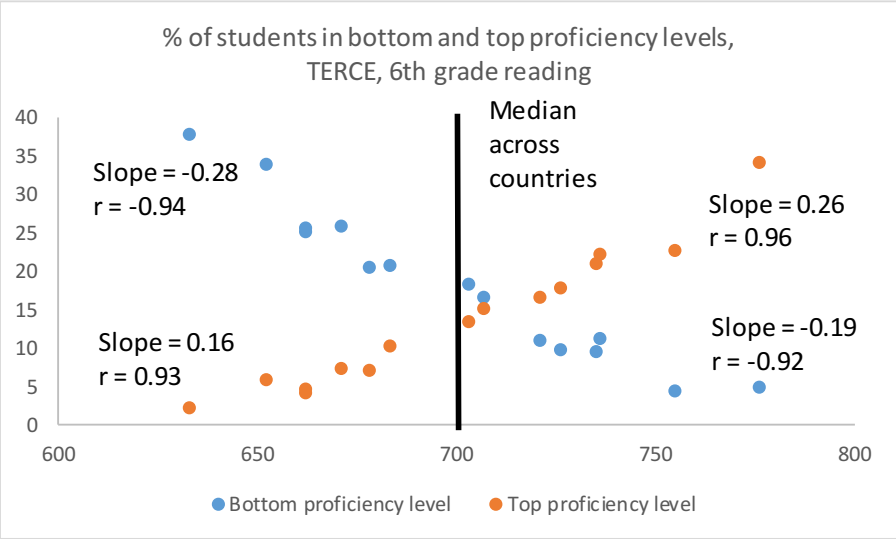

% of students in bottom and top two proficiency levels, PISA 2015 mathematics
Slope = -0.35
r = -0.97
Median across countries
Slope = 0.27
r = 0.91
Slope = 0.06
r = 0.88
Slope = -0.10
r = -0.73
Bottom (1) proficiency level
Top two proficiency levels

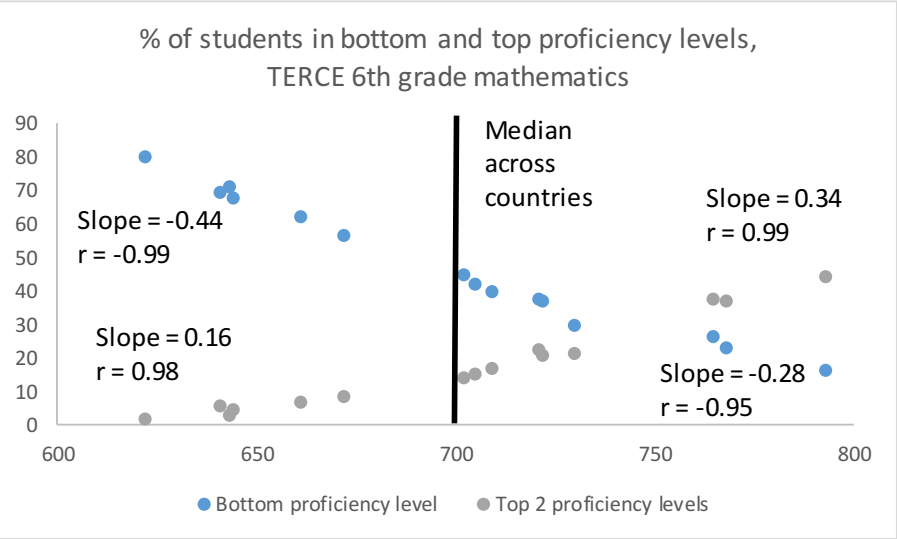

% of students in bottom and top two proficiency levels, PISA 2015 science
Slope = -0.25
r = -0.89
Median across countries
Slope = 0.16
r = 0.84
Slope = 0.04
r = 0.76
Slope = -0.08
r = -0.75
Bottom two proficiency levels
Top two proficiency levels

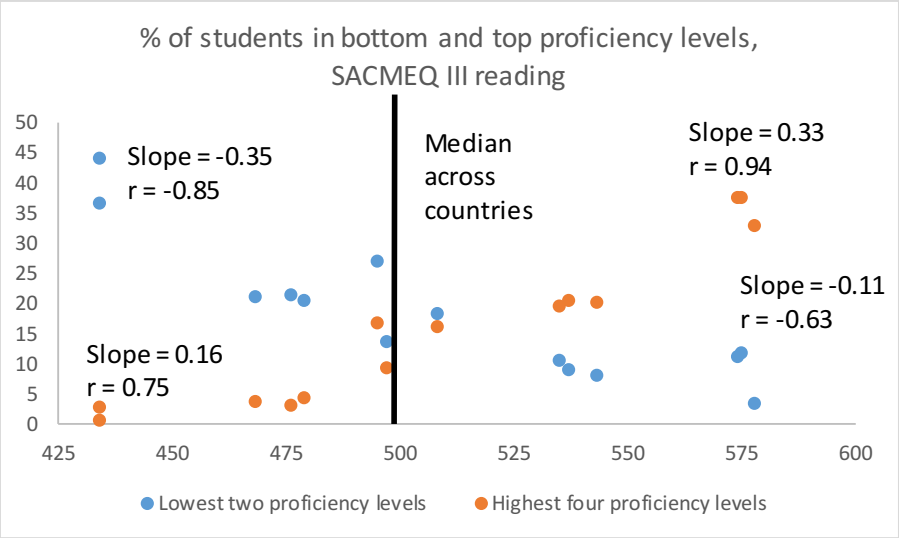

% of students in bottom and top proficiency levels, PIRLS 2011
Slope = -0.35
r = -0.95
Median across countries
Slope = 0.29
r = 0.76
Slope = 0.05
r = 0.72
Slope = -0.10
r = -0.57
Bottom proficiency level
Top proficiency level

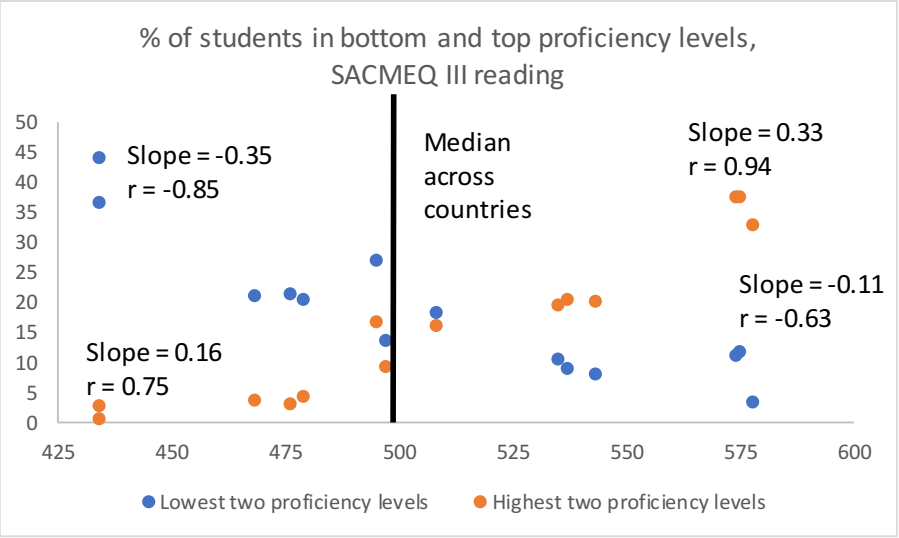

% of students in bottom and top proficiency levels, TERCE, 3rd grade reading
Slope = -0.41
r = -0.99
Median across countries
Slope = 0.29
r = 0.98
Slope = 0.10
r = 0.92
Slope = -0.28
r = -0.98
Bottom proficiency level
Top proficiency level


% of students in bottom and top proficiency levels, TERCE, 3rd grade mathematics
Slope = -0.38
r = -0.99
Median across countries
Slope = 0.18
r = 0.91
Slope = 0.04
r = 0.87
Slope = -0.38
r = -0.91
Bottom proficiency level
Top proficiency level


% of students in bottom and top proficiency levels, TERCE, 6th grade reading
Median across countries
Slope = -0.28
r = -0.94
Slope = 0.26
r = 0.96
Slope = 0.16
r = 0.93
Slope = -0.19
r = -0.92
Bottom proficiency level
Top proficiency level


% of students in bottom and top proficiency levels, TERCE 6th grade mathematics
Median across countries
Slope = -0.44
r = -0.99
Slope = 0.34
r = 0.99
Slope = 0.16
r = 0.98
Slope = -0.28
r = -0.95
Bottom proficiency level
Top 2 proficiency levels


% of students in bottom and top proficiency levels, SACMEQ III reading
50
45
40
35
30
25
20
15
10
5
0
425
450
475
500
525
550
575
600
Slope = -0.35
r = -0.85
Median across countries
Slope = 0.33
r = 0.94
Slope = -0.11
r = -0.63
Slope = 0.16
r = 0.75
Lowest two proficiency levels
Highest four proficiency levels


% of students in bottom and top proficiency levels, SACMEQ III reading
50
45
40
35
30
25
20
15
10
5
0
425
450
475
500
525
550
575
600
Slope = -0.35
r = -0.85
Median across countries
Slope = 0.33
r = 0.94
Slope = -0.11
r = -0.63
Slope = 0.16
r = 0.75
Lowest two proficiency levels
Highest two proficiency levels