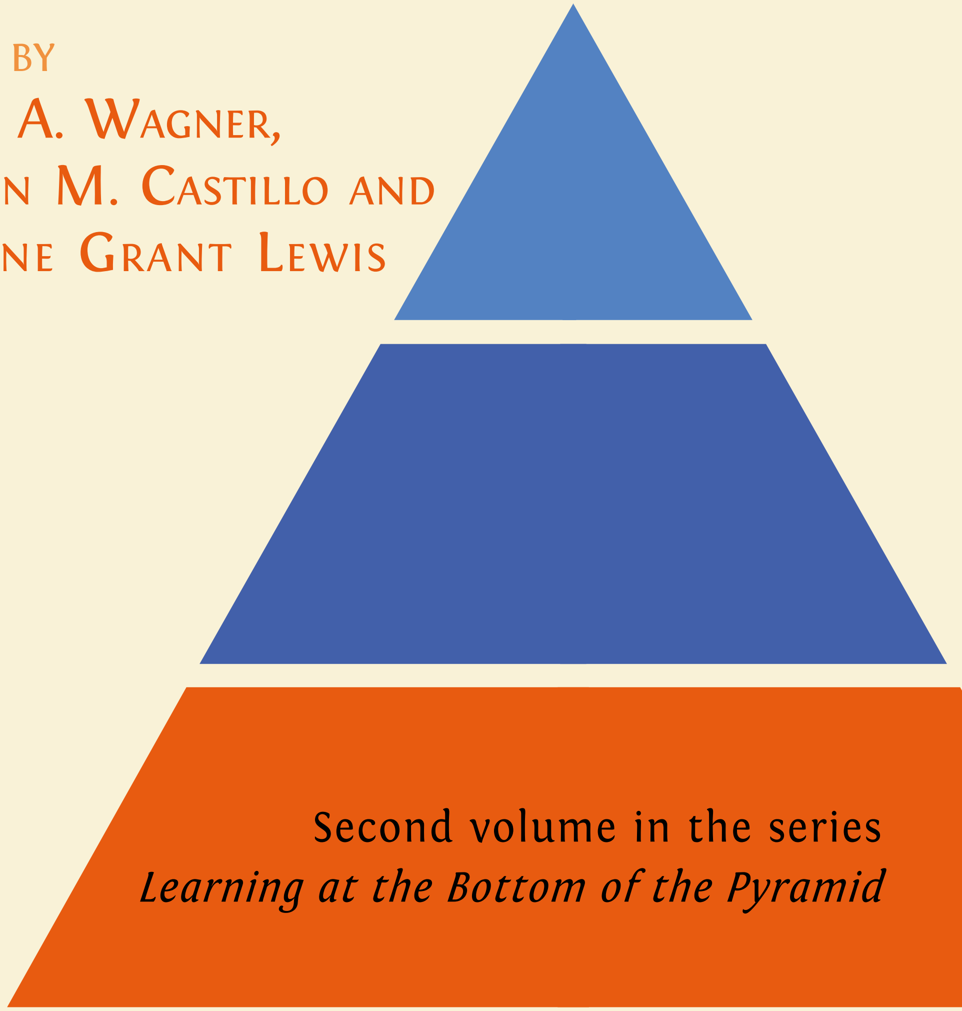
Learning, Marginalization, and Improving the Quality of Education in Low-income Countries

EDITED BY
DANIEL A. WAGNER,
NATHAN M. CASTILLO AND
SUZANNE GRANT LEWIS



Second volume in the series
Learning at the Bottom of the Pyramid

https://www.openbookpublishers.com

Daniel A. Wagner, Nathan M. Castillo and Suzanne Grant Lewis, *Learning, Marginalization, and Improving the Quality of Education in Low-income Countries.* Cambridge, UK: Open Book Publishers, 2022. https://doi.org/10.11647/OBP.0256

ISBN Paperback: 9781800642003
ISBN Hardback: 9781800642010
ISBN Digital (PDF): 9781800642027
ISBN Digital ebook (epub): 9781800642034
ISBN Digital ebook (mobi): 9781800642041
ISBN Digital ebook (XML): 9781800642058
DOI: 10.11647/OBP.0256

Cover design by Anna Gatti.

5. Reducing Inequality in Education Using "Smaller, Quicker, Cheaper" Assessments

Luis Crouch and Timothy S. Slade

Introduction

With the advent of the Sustainable Development Goals (SDGs), and especially SDG4, several important trends have developed within the global education community, particularly in low-income and lower-middle-income countries. Some of the trends pre-date the SDGs, but the SDGs certainly increased focus on them.

First, the world is moving towards the concurrent measurement of access to education *and* learning, not just access (or a learning proxy such as primary-school completion), as was the emphasis under the Millennium Development Goals. In addition, it is moving away from tracking average performance, and is instead focusing on equity and equality. Second, there has been a mushrooming of efforts and data sources that are intended to measure equity and inequality. International and regional assessments continue to grow and adapt by honing their ability to discriminate at the bottom of the learning scale.[1] In particular, there has been an enormous growth in the sorts of "smaller, quicker, cheaper" (SQC) measurements Wagner called for in

[1] E.g., PISA for Development (https://www.oecd.org/pisa/pisa-for-development/) andIEA'sLANA (https://www.iea.nl/publications/presentations/ga56introducing-ieas-lana-developing-countries).

https://doi.org/10.11647/OBP.0256.05

an influential 2003 paper, and further discussed in Wagner's books in 2011 and 2018. Measures like PAL and EGRA have now been used in hundreds of country/language/script contexts.[2] Third, in part because of the dramatic evidence provided by the SQC data, there has been a wave of interest in using the measurements to improve outcomes, precisely as Wagner intended. In one of the latest signs of that interest, the World Bank is seeking to cut the proportion of 10-year-old non-readers in lower-middle-income countries by half, from around 50 percent to around 25 percent, by 2030 (World Bank, 2019). Fourth, there is evidence that countries are beginning to make significant measured progress on some of these fronts, at least at the pilot level (Graham & Kelly, 2019). These efforts have made micro data on pre-and-post treatment-and-control sets available that have typically not been available previously.

This chapter responds to these trends, and will show how they can be potentiated to tackle the issue of learning inequality. It will focus on two issues: (1) whether and how inequality measures can be applied to different sorts of assessment data, especially SQC data, and (2) how different kinds of assessment data, and their corresponding inequality measures, can be used to actually address inequality, along with average performance levels. The focus will be largely empirical, based on data and qualitative observations, and children at the "bottom of the pyramid", as defined in the research by Wagner and others (Wagner, Wolf, & Boruch, 2018). In this chapter, the "bottom of the pyramid" will be represented by the percentage of children who achieve a score of zero on an oral reading fluency exam.

The chapter is structured as follows. After this introduction, a literature review sets out what we could find on the relationship between variations in averages and variations in inequality or (a very different concept) "percent below a learning floor". A subsequent section briefly notes how this chapter differs from that literature, and thus hopes to make an original contribution. The most substantial

[2] People's Action for Learning Network assessments, informally known as Citizen-Led assessments, as at https://palnetwork.org/the-pal-network-case-citizen-led-assessments-to-improve-learning/, Early Grade Reading Assessments as at https://www.sciencedirect.com/science/article/pii/S0738059314001126 and described by Dubeck and Gove (2015).

section of the chapter then uses micro data to test various hypotheses about the measurability of inequality and "percent below a learning floor" and, much more importantly, how these two things co-vary with improvements in the average levels of learning. We do this via reference to two cases (from the same country and modeled on each other) and their corresponding micro data. We give primacy to a set of data from Kenya because of foreknowledge that the data were clean, detailed, and plentiful. The next section describes the substantive literature from the cases. Then, we provide policy implications as to why the data look the way they look, and what the data are telling us about whether and how things change (or don't) for the bottom of the pyramid as averages improve (or don't). Finally, we reprise the finding and provide new research directions in this area.

Literature review

To the best of our knowledge, the application of inequality indices to education—of the sort usually applied by economists to monetary income, wealth, or physical assets such as land—dates to a decade or two after the World Bank first launched significant operations in education. In the early 2000s, Thomas, Wang, and Fan (2001; 2003) published papers calling for the application of the Gini coefficient to educational attainment (years of schooling). At around the same time, the World Bank's World Development Report (2006) was the first major publication by an international agency to provide a systematic compilation of such measures for a relatively large set of countries (28), with a good mix of countries from the low-, middle-, and high-income groupings. The coefficients were based on years of education already received by the adult population, not the current expected years of education. Thus, the concept was analogous, in some sense, to financial or physical wealth.

Strikingly but not surprisingly, these coefficients varied from around 0.2 for countries in Europe and an entity called "C Europe" to around 0.6 for sub-Saharan Africa—the same as what similar tables for income show. While they do not say so explicitly, one can infer from Thomas, Wang, and Fan (2001) that the median value for this education Gini (applied to years of schooling rather than learning outcomes) was about 0.4, which is similar to the current median global income Gini of 0.38. They establish certain interesting facts, such as that the Gini coefficient for years of educational attainment shows quickly-reducing inequality as average years of schooling increase. This is sensible, since individuals are unlikely to pile Ph.D. upon Ph.D. the way they might pile on income.

Both "years of education" and scores on learning assessments could have reasonable upper limits (and certainly do in most international assessments), and therefore the higher the average, the lower the inequality, as high values would be censored from above.[3] However, in those assessments, only a tiny (a few percent at most) of children "top out" at the constructed maximum, and the learning measures used in this chapter do not even have a theoretical or constructed maximum.[4] As noted above, recent calls for this kind of analysis have been associated with the work of Dan Wagner and various colleagues such as in Wagner, Wolf, and Boruch (2018), but their focus has been more on the notion of the "bottom of the pyramid" which is more of a "poverty of learning" concept than a strictly distributionist one.

In addition, other research has asked whether increases in average performance are associated with decreases in inequality or in "percent below a learning floor". These include two existing lines of research (though they typically do not use SQC measurements). One associates differences in cross-sectional data, especially in assessments such as PISA, TIMSS, and PIRLS, to the hypothesized dynamics of increases in average performance; that is, this research looks at whether variations in mean levels of performance are systematically associated with variations in either the *distribution* of performance or the *percent of children below some minimum* ("percent below a learning floor"). Other research looks at these variables as actual changes, over time, in countries that have participated in assessments multiple times.

Papers that compare differences in mean scores cross-sectionally with differences in inequality include Freeman, Machin, and Viarengo

[3] While oral reading fluency, the primary measure discussed in the analysis section, lacks a theoretical upper limit, for all practical purposes it rarely exceeds 200 correct words per minute, especially in the early grades.

[4] The previous five or six sentences borrowed liberally from Crouch and Gustafsson (2018).

(2011), Oppedisano and Turani (2015), Micklewright and Schnepf (2006), Bruckauf and Chzhen (2016), and Sahn and Younger (2007). They also attempt to dig into some of the possible determinants (e.g., Ferreira & Giroux, 2011). A paper by Crouch and Rolleston (2017) as part of the RISE program looked at many of these issues, bringing in evidence from regional learning assessments and special longitudinal studies that measure learning in the same group of children as they grow older (SACMEQ & Young Lives). Little research seems to have explored the long-term changes in the inequality coefficients. One exception was Crouch, Gove, and Gustafsson (2009). Using household surveys from Latin America, we asked about respondents' years of education and their recall of their parents' years of education; the Gini coefficient for years of education improved from 0.58 to 0.36—quite a significant change.

These studies typically focus on one assessment for a specific (sometimes relatively distant) year. Crouch and Gustafsson (2018), on the other hand, systematically look at data from all of the known assessments within a certain period of time and attempt to explore the same issues. Two or three conclusions are relevant here. First, looking at cross-sectional data from most recent assessments, the correlation between differences in average learning levels and differences in the within-country *distribution* of scores is ambiguous: for some, there is a positive association, for some there is a negative association. Somewhat worryingly, the study found that the associations depended on the assessment organization, suggesting that some of the association, positive or negative, could have been due to methodological issues. Second, the paper unambiguously concludes that, in all of the used assessments, differences in average scores between the low scorers and the medium scorers are very strongly associated with reductions in the percent of children below a certain level of proficiency (both in the correlation sense and in the effect-size sense, though these are more or less equivalent in two-factor correlations). The World Bank (2019) calls this percentage the "learning poor", which is analogous with Wagner's concept of the "bottom of the pyramid".[5] Thirdly, the

[5] In this context we will eschew the term "learning poverty" because it connotes, to non-economists, that below that floor there is no learning. But, of course, children do learn even if they are not in school. We know that is not what economists mean,

paper concludes that when time-series *are* available, they confirm the impression from the cross-section analysis: improvements in overall levels are at first strongly associated with reductions in the percentage of children below a learning floor. A methodological point in Crouch and Gustafsson (2018) is that it is hard to study inequality using Item Response Theory (IRT)[6] scores, since they are not a completely natural metric.[7]

The learning metric used in this chapter, oral reading fluency, is also a "natural" metric. However, it is important to clarify that we are not talking about the distribution of knowledge or skill. After all, regardless of whether one is using IRT or classical scores, increases in scores do not necessarily mean the same thing, regardless of the starting value.[8] Studies of the issue by IEA confirm this (Mullis, Martin, & Loveless, 2016), and their results are amplified by Crouch and Gustafsson (2018, p. 29):

> …An analysis by Mullis et al. (2016: 58), … examine(s) the improvements amongst TIMSS Grade 4 countries, between 1995 (but in some cases 2003) and 2015, focusing 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—and often much larger— improvements at the 10th than the 90th percentile. The present analysis…establishes that the movement is towards less 'percent below a learning floor.' Just six SACMEQ countries were considered to have made significant improvements in their national mathematics score

but to prevent communications barriers we will use another term, namely "percent below a learning floor".

[6] The more "modern" technique for scoring learning assessments, which has many advantages, but has one disadvantage in that the scores are not easy to interpret as a "percent correct answer" in a more classical scoring method.

[7] That paper tends to use classical (percent correct) scores, even for the international assessments. It shows that the correlation between classical and IRT scores is so high that one may as well work with the more natural "percent correct" measure.

[8] It is difficult to say, for instance, that a child whose score improved from 50 to 60 improved their knowledge as much as one moving from 60 to 70. Perhaps the questions answered in moving from 60 to 70 were harder. Many economists recognize that the ultimate goal of policy should be utility or happiness, not income. But they tend to talk about the distribution of income, not the distribution of happiness. This is probably by accident, not by wisdom. But it would do for us to talk about the distribution of scores, not of learning.

> between 2000 and 2007. ...Generally, the six SACMEQ countries did see larger reductions at the bottom than gains at the top.

We draw three distinctions in the rest of this chapter. First, we distinguish between *inequality*—especially as a measure of pure dispersion, most often not in association with other putatively causative factors such as gender or socioeconomic status—and *percent below a learning floor* (a measure similar to income poverty). Second, we distinguish between what one may call *pure inequality*—namely "pure" variance or something like it—and variance associated with other factors. *Inequality* and *percent below a learning floor* are clearly not the same thing, and this distinction has proven to be analytically useful in the literature on economic development and, to some degree, education. Only the notion of *percent below a learning floor* is directly relevant to the calls for attention to the issue from the World Bank (2019) and Wagner et al. (2018). Third, the notion of "pure" inequality or "percent below a learning floor", as de-linked from gender, social status, etc., is most relevant to measurement-based standards and practices related to teaching and learning. Measures associated with other factors like gender or income suggest targeting school support based on those factors, whereas measures associated with "pure" inequality suggest targeting support based on learning outcomes.

Points of departure of the present study from the reviewed literature

The research reviewed above uses aggregated data, either reflecting cross-sectional variation or changes over time for the few countries that have participated in given assessments for multiple years. None of it looks at micro data from the same students, or at least the same teachers or classes at different points in time, while controlling for whether a bona fide pedagogical intervention has taken place, if possible.

We focus almost entirely on issues surrounding measurement: do the measures "behave well", do they seem robust, and are they interpretable? We also delve briefly into the pedagogical issues that relate to the changes in measures, but more as a way of showing what one can learn for educational programs and policies, rather than to come to any firm and generalizable conclusions. If the measures do not seem to have any actionable implications for policymakers, they would be of little use.

At the same time, though the evidence from the literature review seems to show that countries can improve their average performance (at least from low to middling levels) by paying attention to the left tail of their score distributions, how precisely they do so is not clear. In this chapter, we focus more on whether micro evidence of improvement in averages also addresses either cognitive inequality or percent below a learning floor (or both). However, the data themselves and some of the qualitative write-ups on improvement efforts do provide some tantalizing early suggestions.

Data and methods

Measures used

Though the most recent calls for a "Gini coefficient" analysis of education and learning inequality (sometimes implicitly) call for that specific measure, it seemed prudent to assess the behavior and utility of several others as well. We chose the following measures for the reasons noted in Table 1.[9]

[9] A good primer on measures of poverty and income inequality is to be found at Haughton and Khandker (2009), available at: https://openknowledge.worldbank.org/bitstream/handle/10986/11985/9780821376133.pdf.

Table 1. Inequality measures.

Measure	Explanation of the measure	Reason for using it
Gini coefficient	A coefficient ranging from 0 to 1, where 0 represents a hypothetical situation where everyone in society owns an equal amount of the good in question (income, wealth, years of education learning as measured by a test score), and 1 represents a situation where one individual in society owns all of the good in question. Note that this is a measure of relative inequality. If, for instance, everyone’s incomes or test scores were to increase by the same absolute amount, the Gini coefficient would decrease, even if the absolute distance between the income (score) of the highest earner (scorer) and the lowest is the same as before. In the rest of this chapter, whenever we use the term “Gini coefficient” we will mean the Gini coefficient as applied to just one measure of learning, namely oral reading fluency, unless we explicitly say otherwise in order to refer to income or wealth or some other underlying concept.	It is best known to both economists and non-economists, and has a one-to-one correspondence with a well-known graphical interface, the Lorenz curve.[10] The Lorenz curve can be used to visualize where in the distribution the inequality comes from, and can be linked to concepts of absolute poverty or “percent below a learning floor”. (These concepts are used in the chapter, and the Lorenz curves are shown.) It has also been the most used measure in education (but mostly for years of schooling).

[10] For a comparison between the Lorenz curves for the learning outcomes shown here and some income distributions from very equal and very unequal countries, see the section called “Lorenz curves for learning and for income” below.

Measure	**Explanation of the measure**	**Reason for using it**
Coefficient of variation	Standard deviation over the mean. Zero lower bound, no theoretical upper bound. As with the Gini coefficient, an equal absolute increase in income or test scores, for everyone, would make the measure decrease.	It is easy to calculate even with common software such as Excel. No specialized substantive or computational knowledge is required.
Ratio of Px to Py, typically 90th to 10th or 75th to 25th	Ratio of a measure (e.g., income, score on a test) characteristic of the person at the x^{th} percentile of a distribution compared to the same characteristic at the y^{th} percentile of the distribution. Lower bound of 1, no upper bound. A relative measure in the same way as the others.	It is intuitively appealing and often used, analogous to the popular economic and political literature around "percent of wealth possessed by the one-percenters".
Generalized Entropy [GE()] index with	Generalized entropy indices are a class of income inequality measures. The parameter governs the weight given to the distances between two incomes along the distribution; smaller values of increase sensitivity to changes at the lower end, while larger values of increase sensitivity to changes at the upper end.[11] GE(0) and GE(1) are undefined for incomes of 0, and are therefore an ill fit for our data, which feature large proportions of zero scores. We use GE(2).	Though they are the least known outside the economics profession, they have a few advantages. One is that there are various measures and they can be relatively more or less sensitive to inequality in certain portions of the distribution (e.g., more sensitive to the inequality amongst the poor than amongst the rich). A second is that they are easily decomposable in a manner similar to analysis of variance: e.g., inequality within schools and between schools, adding up to total inequality.

[11] World Bank Institute. (2005). *Introduction to Poverty Analysis*.

Measure	Explanation of the measure	Reason for using it
Percent scoring zero	This is a measure akin to what the World Bank is calling "learning poverty", what we are calling *percent below a learning floor*. It is similar to income poverty, which is measured as the percent of people below a certain income threshold. In our case we have chosen zero—the inability to read a single word—as a dramatic cut point, and one whose improvement would seem to be able to easily draw attention and bureaucratic effort.[12]	Very easy to interpret and possibly act upon.

[12] For the notion and use of "learning poverty" see World Bank (2019), where they define as "learning poor" any child not in school or not able to read and understand a simple paragraph by age 10. This is a more advanced age and level than the specific definition of "percent below a floor" we use in this chapter, which is the percent of children unable to read even one word.

Data and methods, PRIMR and Tusome in Kenya

Over the last decade, the Early Grade Reading Assessment (EGRA) has been used in more than 70 countries and 120 languages to estimate the reading abilities of primary school learners, using a variety of reading and pre-reading metrics (RTI International, 2015). EGRA is comprised of tasks designed to measure skills such as phonological awareness, decoding, listening comprehension, and others. But policymakers frequently focus on learners' results on the oral reading fluency (ORF) metric, as it is the closest analogue to the common "educated layperson's" understanding of what "being able to read" means—independent reading of narrative text.[13]

From the earliest days of EGRA, and using various classical analyses of EGRA results, it was the skill that had the highest item-test and item-rest correlation, and is the one that weighs most heavily in factor analyses attempting to discern whether there is a latent construct that can be called "early grade reading skill". These correlations or associations are all highly statistically significant and, substantively, follow the patterns one would hope (e.g., the principal-components analysis has a big first-factor weight, a big drop-off between the first and second factor, and the sub-skills load reasonably evenly onto that first factor). For the two EGRA applications in PRIMR and Tusome, described below, the Cronbach's alpha measures 0.81 and 0.86 respectively.

In the data that comprise the source for this chapter, the child is presented with a simple story of approximately 60 words in length and is asked to read aloud as much of it as they can within one minute. If the child is unable to complete the text within the minute, the exercise is stopped and the last word they attempted to read is noted. If the child reads the entirety of the text before the minute elapses, the assessor stops the timer and notes the amount of time remaining. In either case, the assessor tracks the child's progress and marks any words that the child reads incorrectly.

ORF is reported in *correct words per minute* (cwpm). For children who do not complete the passage, their cwpm score is simply the number of words they read correctly. For children who complete the passage with time remaining, the number of words they read correctly is transformed into a cwpm score according to the formula as follows:

[13] Silent reading skill as an addendum to EGRA tasks is being piloted.

$$items\_per\_minute = \frac{items\_correct}{(time\_for\_task — time\_remaining) * 60}$$

The result is a continuous measure bounded from below at 0.[14] Children who are unable to read a single word correctly obtain a *zero score*. The inequality analyses presented in this chapter depend upon a continuous measure and are therefore appropriate for use with EGRA data. Note that there are many other tasks in a typical EGRA application, from more phonemic ones to others aimed at comprehension.

The ORF data used in these analyses were collected under the United States Agency for International Development (USAID) Primary Math and Reading Initiative (PRIMR) and Tusome Early Grade Reading Activity ("Tusome").

PRIMR was a partnership between USAID and Kenya's Ministry of Education, Science, and Technology (MoEST), meant to identify mechanisms to improve reading outcomes in Kiswahili and English (RTI International, 2014). It was implemented from 2012–2014 in 547 government primary schools and low-cost private schools (LCPS) in four Kenyan counties: Nairobi, Murang'a, Kiambu, and Nakuru. PRIMR used a three-cohort design: Cohort 1 received the intervention from 2012–2013, Cohort 2 from 2013–2014, and Cohort 3 was retained as a control until after the endline data collection had concluded. Kiswahili and English reading outcomes among Grade 1 and Grade 2 children were assessed using EGRA at baseline, midline, and endline, and using comparison groups.[15]

Tusome was a partnership between USAID and Kenya's MoEST that brought the most promising interventions from PRIMR to *all* public primary schools in the country and 1,500 LCPS. While the intervention was ultimately extended to Grade 3, the external impact evaluation only assessed the Kiswahili and English reading performance of Grade

[14] There is no fixed theoretical maximum, as it depends on the total items in the task and the time allotted. Practically speaking, it is extremely rare to find ORF scores exceeding 200 cwpm. Adults with many years of education and who read for a living, but without training in speed reading, can read (aloud) about 200–220 cwpm at an unforced pace.

[15] The current analyses do not include data from PRIMR's "ICT Pilot" in Kisumu County.

1 and Grade 2 children. Given the universality of Tusome, there was no control group. The analyses shown in this chapter incorporate data from Tusome's baseline and midline EGRAs, conducted in July 2015 and September through October 2017 respectively. The data were collected from 204 schools (of which 174 were public and 30 LCPS) according to a three-stage cluster sampling approach designed to yield nationally representative estimates.

Results

Results are presented separately for PRIMR and Tusome. Initial analyses focused on whether specific measures of inequality could be computed, and if so, whether they appeared to behave in ways that would be consistent with theory. Separate analyses for Kiswahili and English are presented for key subpopulations defined by grade (Grade 1 vs. Grade 2) and round of assessment (baseline, midline, or endline). For the PRIMR data, additional breakdowns are provided by cohort, which capture treatment status and duration of intervention (see Table 2). As Tusome was a nationwide intervention in all public primary schools, the results are from treatment schools. We focus exclusively on the oral reading fluency score, as it is our best available proxy for reading ability.[16]

Basic results, behavior, and interpretation of the measures

The following section first shows basic results that help us decide whether the measures are "well-behaved".[17] We then draw out some of the substantive interpretations. Table 2 and Table 3 below report estimates

[16] Reading comprehension measures better represent the actual goal of reading. However, EGRA's reading comprehension measures have very few (five) items and are categorical in nature, making them ill-suited for this analysis.

[17] We do not mean "well-behaved" in a particularly rigorous manner. In general we are looking for "good behavior" in the sense of ratios that do not become infinite or undefined, indicators that move more or less with each other (so that the Gini coefficient does not decrease significantly while the coefficient of variation goes up, say), or numerical results whose directional movement ends up making intuitive sense and lines up with some reasonable substantive narrative that fits the observed facts. In some sense, this notion of "good behavior" is meant to answer questions such as "does it seem possible to measure learning inequality in these ways", and "does that measurement make sense for the context and situations noted?"

for several inequality measures: the Gini coefficient, the Generalized Entropy Index with $(GE(2))$, the ratio of the 90th percentile score to the 10th percentile score (*ratio_p90p10*), the ratio of the 75th percentile score to the 25th percentile score (*ratio p75p25*), and percent scoring zero (*pct_zero*). The tables also include estimates for the mean fluency and associated coefficient of variation (CV) for each subpopulation. The mean is presented not as a measure of inequality (which it is not, of course) but because without it, it is harder to interpret the measures of inequality that are presented (see Table 2).

In terms of being "well-behaved", several patterns emerge.

1. **Ratio of Px to Py**. In nearly every subpopulation, *ratio_p90p10* cannot be calculated because more than 10 percent of the children assessed recorded a score of zero. While *ratio_p70p25* can be calculated more frequently, it is available for fewer than 50 percent of the subpopulations, and far less frequently for Kiswahili than English. We know from the work of other colleagues that this ratio, applied to other datasets, also tends to break down (e.g., Dowd, 2018). It may be that, in spite of the intuitiveness of the ratio and its easy use in income and wealth analysis, it is the least usable of all the measures assessed. In fact, the measure is so ill-behaved that we find it difficult to say anything substantive based on it. However, for the sake of illustration, in some 30 rich countries, Gromada et al. (2018) found a median ratio of only 1.41 (for p90/p10) for education (reading, primary school), considerably lower than the 5.0 (estimating across all our measures) that we observe.[18] In a variety of US household surveys from the 1990s, Hao and Naiman (2010) show this ratio to be around 25 on average, for income.

2. **Gini coefficient**. The Gini coefficient for learning seems to consistently behave well. The values observed are in line with what one observes from comparative studies or simple

[18] The 1.38 for the Gromada et al. (2018) study is our interpretation. Given that they use a metric without a valid zero, we projected "zero" as our projected score of the child at 1st percentile, subtracted that from scores at all the key percentiles, and then calculated the p75/p25 ratio.

Table 2. Range of inequality measure results, PRIMR.

Language	Grade	Cohort	Round	mean	CV	ratio_p90p10	ratio_p75p25	Gini	GE(2)	pct_zero
Kiswahili	Gr 1	1 (Full Tx)	Baseline	4.8	15.1	●	●	0.826	2.29	71.0
			Midline†	21.6	4.5	●	11.3	0.484	0.37	23.0
			Endline†	19.1	5.6	●	●	0.527	0.49	28.4
		2 (Delayed Tx)	Baseline	4.9	7.7	●	●	0.771	1.51	64.4
			Midline	19.6	3.8	●	5.6	0.451	0.32	17.8
			Endline†	20.8	4.2	●	3.8	0.455	0.34	22.3
		3 (Control)	Baseline	3.3	9.4	●	●	0.848	2.60	73.8
			Midline	15.4	4.8	●	●	0.522	0.45	28.6
			Endline	13.4	5.7	●	●	0.517	0.44	28.7
English		1 (Full Tx)	Baseline	6.8	18.4	●	●	0.823	2.45	66.3
			Midline†	30.6	5.9	●	16.6	0.534	0.48	24.0
			Endline†	29.9	5.6	●	12.8	0.543	0.51	22.7
		2 (Delayed Tx)	Baseline	7.5	8.1	●	●	0.750	1.46	54.0
			Midline	29.4	5.0	●	7.7	0.505	0.43	20.1
			Endline†	33.7	4.1	●	4.4	0.469	0.36	18.3
		3 (Control)	Baseline	4.4	10.0	●	●	0.852	2.75	72.4
			Midline	19.6	6.3	●	●	0.620	0.73	31.4
			Endline	20.1	7.0	●	36.0	0.569	0.57	26.3
Kiswahili	Gr 2	1 (Full Tx)	Baseline	17.0	7.7	●	●	0.542	0.49	33.6
			Midline†	32.4	3.5	●	●	0.369	0.22	10.7
			Endline†	32.0	3.4	●	●	0.362	0.21	10.4
		2 (Delayed Tx)	Baseline	19.7	4.1	●	●	0.479	0.36	25.5
			Midline	32.5	3.5	●	14.2	0.344	0.19	9.1
			Endline†	34.4	3.2	●	5.5	0.309	0.16	7.2
		3 (Control)	Baseline	15.1	5.1	●	●	0.578	0.58	36.5
			Midline	26.6	3.7	●	●	0.358	0.20	12.4
			Endline	26.9	3.7	●	●	0.352	019	10.5
English		1 (Full Tx)	Baseline	25.3	9.5	●	●	0.553	0.52	30.1
			Midline†	48.6	4.4	●	3.7	0.424	0.29	11.5
			Endline†	56.0	3.2	26.6	2.9	0.382	0.23	9.3
		2 (Delayed Tx)	Baseline	28.3	4.6	●	9.0	0.496	0.40	19.0
			Midline	53.7	3.8	16.7	2.6	0.372	0.22	8.6
			Endline†	60.6	3.3	7.1	2.2	0.318	0.16	5.5
		3 (Control)	Baseline	18.8	5.6	●	●	0.626	0.73	33.0
			Midline	36.8	4.3	●	4.1	0.439	0.32	10.7
			Endline	42.8	4.3	●	3.0	0.402	0.26	13.1

† = had received the intervention in the period preceding the assessment.

● = could not be calculated.

databases of income and wealth inequality.[19] In recent years, according to World Bank estimates, the most income-unequal countries (taking together all countries' unique measurement points in the last five years), were South Africa and Namibia, with income inequality Ginis around 0.6, and the most equal were some of the ex-Soviet countries such as Ukraine and Belarus, with Ginis for income around 0.25, similar to many of the Nordic countries. Below we can see the Ginis for learning consistently (but not invariably) decrease from baseline to endline, and are similar to the income Ginis for unequal societies. At a glance, the Gini appears to be roughly comparable across languages, and tends to be smaller in Grade 2 than in Grade 1. This is also intuitive: we would expect Grade 2 scores to have less variation, as longer exposure to the school system begins to smooth out the effects of household-level factors.

3. **Coefficient of variation**. This indicator also behaves well. The correlation coefficient between the CV and the Gini is 0.84 (across all cohorts); these two measures of inequality move together well and tell more or less the same story. It does have the disadvantage of not having a theoretical upper bound, and it can be more influenced by outliers than the Gini coefficient.

4. **Percent scoring zero**. The "percent scoring zero" also behaves well. Analysts working on fluency have been analyzing these data for some time, and do not report major issues with this measure, so one would expect this.[20] Note that this indicator is more akin to the concept of "percent below a learning floor" than inequality and, as noted in Crouch and Rolleston (2017) and Crouch and Gustafsson (2018), this indicator may matter more. The correlation with the Gini coefficient across the observed data points is 0.97. As will be noted below, this measure influences the Gini coefficient in a very understandable and reasonable way.

[19] The data referenced here are from a download of the World Bank's World Development Indicators, that can be found at https://datacatalog.worldbank.org/dataset/world-development-indicators.

[20] See examples from many countries at https://earlygradereadingbarometer.org/.

5. **GE(2) index.** The GE(2) index appears to behave well, too. Values generally but not uniformly decrease over time, both from baseline to endline and from Grade 1 to Grade 2. Any differences between the GE(2) values for English vs. Kiswahili within a given cohort and round of assessment appear slight and may not be meaningful. For all subpopulations in both languages, the GE(2) was substantially reduced by midline relative to the baseline; changes between midline and endline were comparatively modest, and sometimes reversed course. In general the GE(2) measures seem more sensitive than the Gini coefficient: they show bigger changes between baseline and midline. Whether that would continue to be the case with other datasets is unknown for now.

Table 3. Range of inequality measure results, Tusome.

Language	Grade	Round	mean	CV	ratio_p90p10	ratio_p75p25	Gini	GE(2)	pct_zero
Kiswahili	Gr 1	Baseline	4.9	11.8	●	●	0.819	2.02	69.9
		Midline	12.2	6.1	●	●	0.634	0.75	43.0
	Gr 2	Baseline	13.5	6.8	●	●	0.617	0.68	43.3
		Midline	24.5	4.2	●	2.73	0.401	0.25	18.4
English	Gr 1	Baseline	10.6	10.6	●	●	0.741	1.36	52.8
		Midline	22.3	6.2	●	12.3	0.572	0.58	20.9
	Gr 2	Baseline	23.8	7.3	●	●	0.615	0.68	37.9
		Midline	43.6	4.7	32.6	3.0	0.397	0.24	10.7

● = could not be calculated.

The general patterns we observed in the PRIMR data—which contained more schools, but were from a narrower representative sample—are reflected in the Tusome data as well. The *ratio_p90p10* can rarely be calculated, and while *ratio_p75p25* is available slightly more often, it is not consistently so. The Gini coefficient appears roughly comparable across languages and tends to diminish both over time, both from baseline to midline and from Grade 1 to Grade 2. The coefficient of variation in both languages narrows from baseline to midline and from Grade 1 to Grade 2, while the mean scores increase. Likewise, the percent scoring zero and GE(2) diminish substantially from baseline to midline and Grade 1 to Grade 2 for both languages.

Selected graphical analyses and interpretation

The tabulations above suggest the *ratio_p90p10* and *ratio_p75p25* are unlikely to present fruitful avenues of exploration and the GE(2) is unfamiliar to non-economists. So, we set them aside in favor of further exploring the Gini coefficient, the CV, and the percent scoring zero.

Fig. 1. Comparison of Gini measures at t0 and t1 to chart improvement. presents the Gini coefficients separately for each language for various subpopulations at t_0 and t_1, where t_0 is a reference point and t_1 represents a subsequent round of data collection. For Tusome, all t_0 are baseline and all t_1 are midline. For PRIMR, a given t_0–t_1 pairing may be any of baseline–midline, baseline–endline, or midline–endline.[21]

The reference line is the line of equality (but only in the definitional, mathematical sense, not the Lorenz curves sense—see below). If the Gini coefficient for a given comparison were the same at both time periods, the dot would be plotted on the line of equality. If it has diminished from t_0 to t_1, the dot will move farther below the line. If it has increased from t_0 to t_1, the dot will move up toward the line.

Fig. 1. Comparison of Gini measures at t0 and t1 to chart improvement. also shows the Gini coefficient for each subpopulation at baseline, and at the next period. Perhaps the most interesting and immediately visible point is that the results are strongly patterned. One does not have to know which subpopulation each dot represents to see the pattern.

[21] We acknowledge that this approach results in some duplication of data. Removing internal points (e.g., midline for PRIMR Cohorts 1 and 3) would eliminate some of that duplication, but risk introducing either varying durations for the t0–t1 period (baseline-endline for Cohorts 1 and 3, but midline-endline for Cohort 2) or preserving the durations but muddying the treatment/control status (as for PRIMR Cohort 2, which was a control group from baseline to midline before receiving treatment between midline and endline).

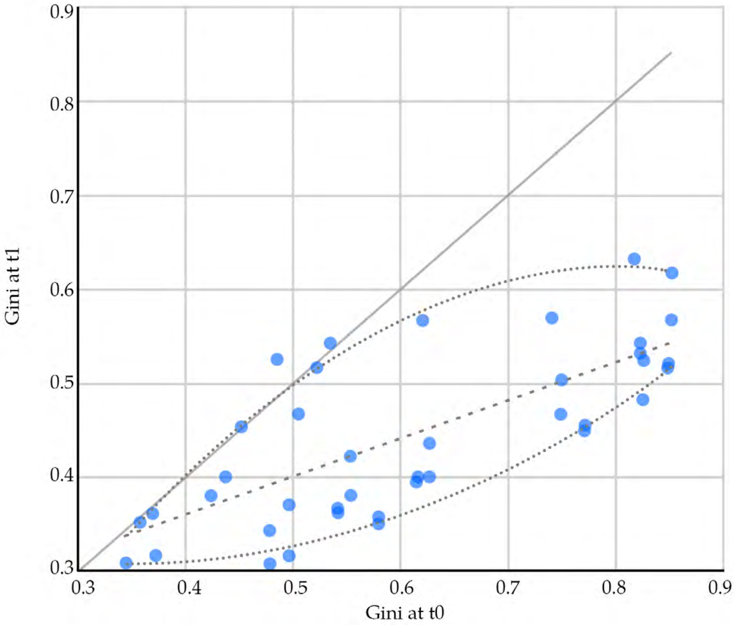


Fig. 1. Comparison of Gini measures at t_0 and t_1 to chart improvement.

The gray line on the graph represents the 45-degree line of equality as noted above. The fact that almost all points are below it tells us that, in almost all cases, the Gini improved.

In addition to the 45-degree line of equality, we have overlaid three other lines. The first one (middle line, dashed) is a simple linear regression. It obviously has a slope less than 1, as can be seen when comparing it to the 45-degree line of equality. The other two dotted lines are "quantile regressions" that provide the best (non-linear, in this case) fit through the scatter at the 15th and 85th percentile on the vertical axis for every point on the horizontal axis.

We interpret these overlays as follows. First, using the 45-degree line of equality, we can see that, as noted, the Gini coefficient nearly always either diminishes from t_0 to t_1 or stays the same, but rarely gets worse. The interventions nearly always improve equality.

Second, we truncated both the horizontal and vertical axes at 0.3, which is more or less the smallest value for both the t_1 and t_0 Ginis. This just helps us concentrate better on the more meaningful parts of the graphical analysis. Note that the regression through the points has a slope much smaller than 1. This is telling us that at low values of inequality in the baseline (t_0), it is harder to further reduce the value by t_1—at around 0.3 for t_0, we end up at about 0.3 for t_1. But at 0.8 for t_0, things have improved all the way to 0.53 or so for t_1. This makes sense given the concept of diminishing returns—but it is interesting how strong it is. Thirdly and finally, the "buttonhole" shape created by the lines at the 15th and 85th percentiles means that at both extremes (low-starting and high-starting inequality, or a low- and high-starting Gini at t_0), the shift in Gini by t_1 is more *predictable*—not bigger or smaller (we have already noted that the bigger it is at baseline the more it improves), but rather, we are saying that for low- or high- starting Ginis the change is more *reliable*. With Ginis of 0.3 and 0.8 at t_0, the range of improvement is about 0.05 and .15 respectively, but with a Gini of 0.6 at t_0, the range of improvement is about 0.22, a positive outcome.[22]

In Fig. 2. Non-readers at t0 and t1., the percentage of non-readers, or "percent scoring zero", shows the same pattern as the Gini measures. Comparison of the dots to the grey line of equality shows that the percentage of non-readers nearly always improves, and the dashed regression line shows that the worse the value is at baseline (i.e., the more children reading at 0), the more it improves—by a lot.

Improvements in mean reading fluency in Fig. 3. Mean reading fluency at t0 and t1. show the same pattern as the inequality measures, but, as one would hope, in reverse: almost all the points are above the 45-degree line, showing that skills almost always improved. The dashed regression line through the observations does not have a slope very different from the 45-degree line. It is slightly flatter than the line of equality, which suggests that at low levels of fluency it is slightly easier to make gains—again as one would hope, if not expect, given all the foregoing (see Figure 3).

[22] It is possible that we are seeing some regression towards the mean in these results. However, it seems doubtful, given that the observations are not related to schools, but to skill types and levels. Also, there is actually no regression: "good" values in one period do not regress. It also seems less likely with measures of inequality than with point measures. A simple principle more likely at work is something akin to the law of diminishing returns: when one starts at a relatively good place, it may be harder to move forward. On the other hand, while laws of diminishing return make a lot of sense in fairly simple production processes (the returns of 100 kg more fertilizer on a field of corn are not the same at high existing levels of fertilization than at low), it is not as obvious that they would operate similarly in complex social and managerial situations, such as school improvement.

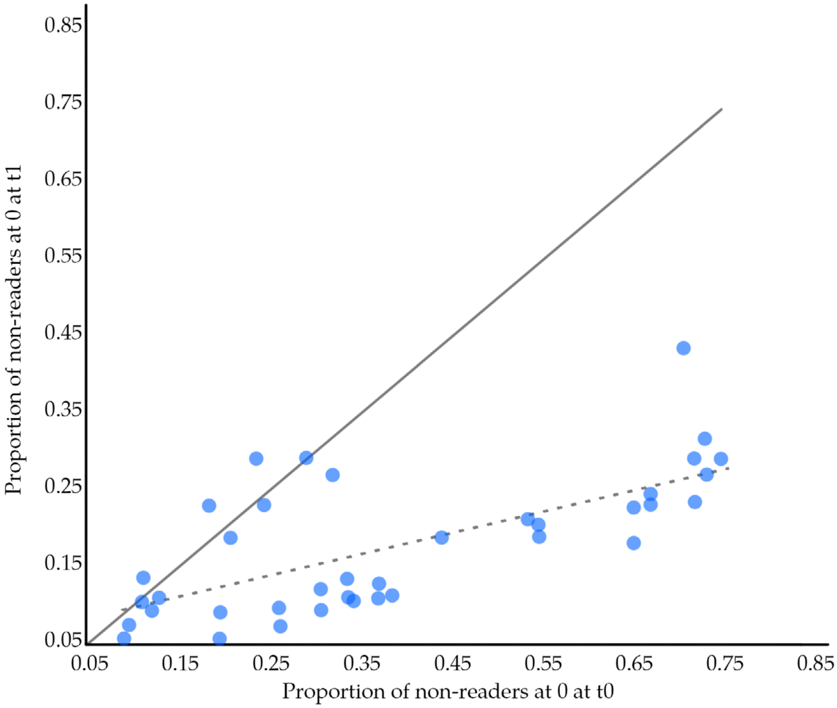


Fig. 2. Non-readers at t_0 and t_1.

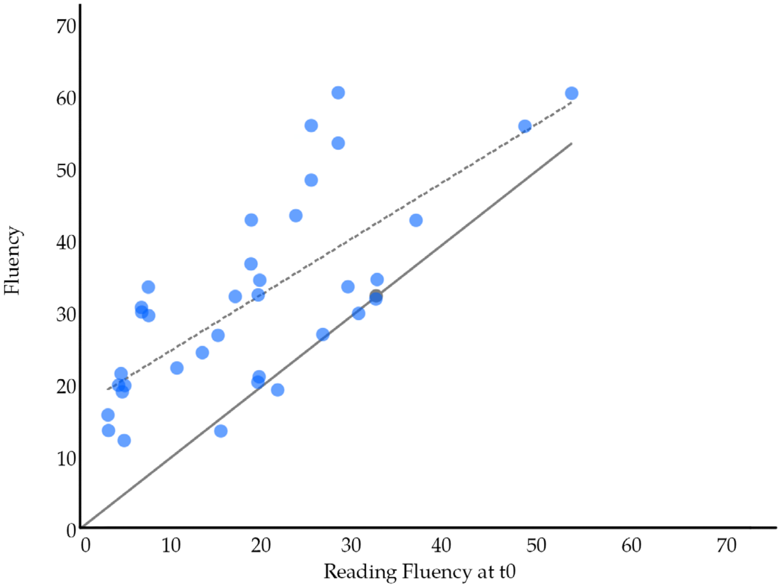


Fig. 3. Mean reading fluency at t_0 and $t_{1.}$

The graphical and statistical analysis done thus far shows that the measures of inequality (Gini and CV) or percent below a learning floor (percent scoring zero) and the means for reading fluency all behave as expected, and in ways that are eminently interpretable—at least in data from a couple of successful and related projects.

But the most interesting and important question is whether the improvements in means from t_0 to t_1, cohort by cohort (as shown in Table 1 and Table 2), were correlated with reductions in inequality for the same cohorts between t_0 and t_1. Fig. 4. Changes in the mean and changes in inequality. shows the correlation between improvements in mean reading fluency and reductions in the Gini of reading fluency (and not for any of the other inequality measures). The dark dashed line is the standard regression and the dotted lines are the quantile regressions. First, the correlation, at -0.65, is strong. The slope is also fairly strong: with fluency improvements of around 10, the Gini improves by -0.15, but with fluency improvements around 25, the Gini is improved by about -0.25. This is notable, given that the starting Ginis were pretty high. So, we can strongly conclude that in this case, the bigger the improvement in the means, the greater the reduction in inequality in oral reading fluency. We believe that this is a very important result.

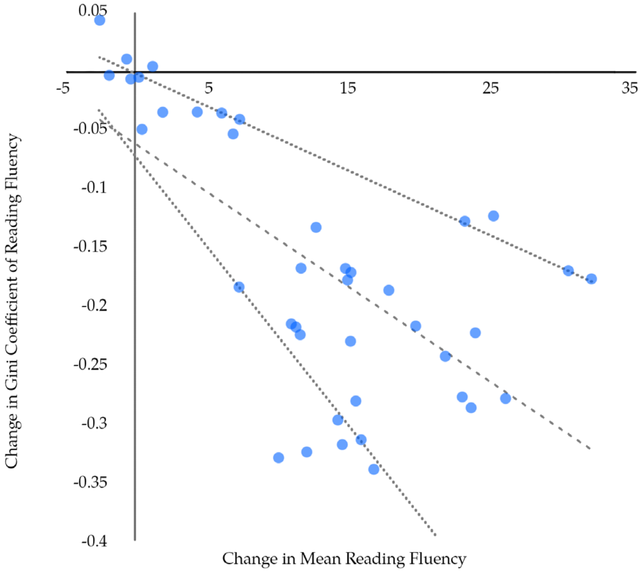


Fig. 4. Changes in the mean and changes in inequality.

Graphical link between the Gini measures and the percent scoring zero measures

One advantage of the Gini coefficient, aside from it being a widely used measure of inequality, is that it has a graphical equivalent in the Lorenz curve. It also seems to work well with the Gini coefficient for learning, with the measures as used in this chapter. In this section we first explore what the Lorenz curves tell us, then explain a link between them (and the Ginis they represent), and the "percent scoring zero" measure of percent below a learning floor.

Fig. 5. Shifts in Lorenz curves in response to successful interventions. provides another way of reasoning about the distribution of oral reading fluency within a subpopulation. It displays the Lorenz curves for Grades 1 and 2 with respect to oral reading fluency, assessed in English and Kiswahili. In this instance, a point on the Lorenz curve can be interpreted as "the bottom X percent of children possess Y percent of the total fluency". The reference line is the line of equality: if fluency skills were equally distributed among the population, for instance, the bottom 20 percent of the children would have 20 percent of the fluency skill. The gap between the reference line and the actual curve represents the inequality of distribution; as the actual curve approaches the line of equality, fluency is distributed more equally, and the gap between the fluency *haves* and *have nots* is closing. The more bowed towards the right-hand bottom corner, the more inequality the curve represents. The dashed lines around each Lorenz curve are the confidence intervals for the curves.[23] The link between the Lorenz curves and the Gini coefficient is simple: the Gini represents double the area between the line of equality and the Lorenz curve. It is important to keep this in mind when using the Lorenz curves to analyze the inequality.

Consider the top left panel in the figure, representing the distribution of English-language ORF among Grade 1 children assessed under PRIMR. At baseline (represented by the yellow line), the bottom 80 percent of children together represent roughly 20 percent of the total English-language ORF observed. As children at the lower end of the

[23] We do not dwell further on the issue of statistical significance as it is generally extremely high, and it is not the issue of interest—but it is good to just establish that the differences are generally very significant.

skill distribution improve their performance, the inequality diminishes and the gap between the *haves* and *have nots* begins to close: by endline (represented by the red line), the same 20 percent of English-language fluency is held by "only" the bottom 60 percent of the children. The bottom 80 percent of children, who formerly held only 20 percent of the fluency, now hold nearly 50 percent of it.[24]

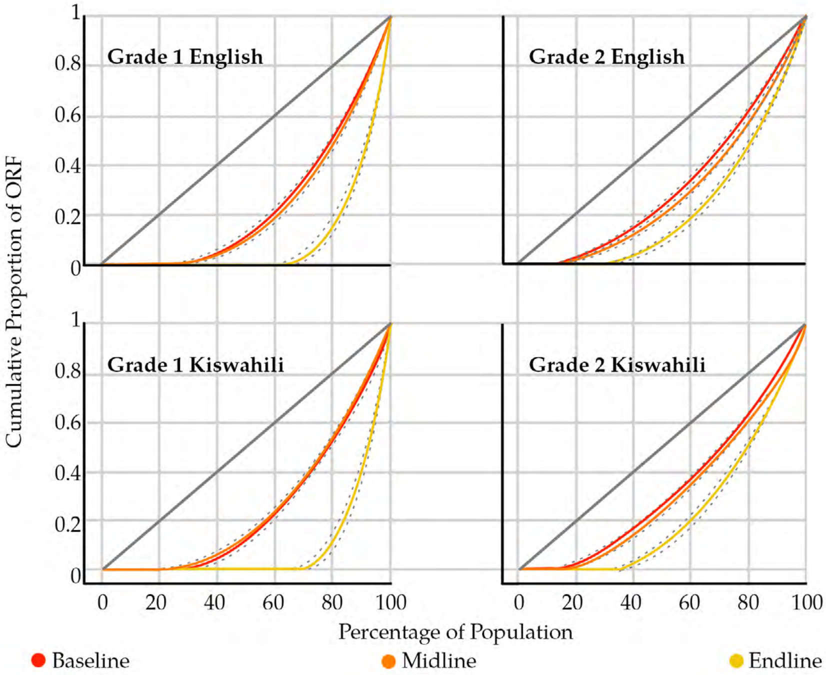


Fig. 5. Shifts in Lorenz curves in response to successful interventions.

Note that the interesting comparisons are not only across the grades and languages, but also between the baselines, midlines, and endlines. Fluency is unequally distributed in all cases. However, it is more unequally distributed in Grade 1 than in Grade 2, as would be expected due to unequal access to resources at the household level. While Kiswahili fluency is slightly more unequally distributed than English in

[24] Recall that these Lorenz curves are analogizing from income, thus the percent of fluency is "held" by a given percentage of the children, exactly as one would say for income or wealth.

Grade 1 at baseline, by midline and endline (and throughout Grade 2) it is *less* unequally distributed.

In each subpopulation, an initially enormous gap between the *haves* and *have nots* has been substantially reduced by endline. In the case being analyzed in this chapter, the vast majority of that gap is closed between baseline and midline, with very little change in any subpopulation between midline and endline. This is especially interesting because the time elapsed from baseline to midline (10 months) and from midline to endline (12 months) is approximately the same. This aligns with what would be expected in the context of an intervention like PRIMR, which explicitly prioritized the teaching of basic literacy skills (such as letter recognition and decoding) before addressing higher-order literacy skills (such as reading with automaticity for comprehension).

We can now explore the link between the Lorenz curves and the "percent scoring zero"—a measure of percent below a learning floor. From the graphics, it is clear that these Lorenz curves all have a flat portion at the left, essentially the same as the horizontal axis, and then bump up a bit further to the right as the curve departs from the horizontal axis. The length of the flat line to the left of the bump represents the percentage of the population "scoring zero". The interpretation is clear: since the children to the left of that bump read at zero, they will cumulatively "possess" zero percent of the cumulative proportion or distribution of fluency—the variable represented on the vertical axis—thus the line is flat and is the same as the horizontal axis. It is particularly interesting that it is the shifting of that bump that accounts for much of the decrease in how "bowed out" the curve is towards the right-hand lower corner of the graphics. That is to say, the reduction in percent below a learning floor (percent scoring zero) accounts for a great deal of the improvement in inequality.[25]

Policy implications

This chapter focused on answering several simple questions: do inequality measures typically used in socioeconomic analysis work for learning, and do they detect levels and changes (in response to

[25] This is evident visually, and we do not quantify it here, but it would be possible to do so.

interventions) that are interpretable and meaningful? Do various measures of inequality (such as the Gini coefficient) and measures of percent-below-a-minimum correlate and reinforce each other, again, in ways that are interpretable and meaningful? The results of the data analysis carried out for the chapter strongly suggest that the answer to these questions is "yes".

But what accounts for the changes we observed? From a policy or pedagogical standpoint, how do the indicators help us formulate actions that could improve performance at the bottom of the pyramid? The data from the implementations and pilots reported here were not designed explicitly to deal with these questions. However, the results are strongly suggestive.

There are strong indications as to the possible causal mechanisms in some of the scholarly literature coming from PRIMR and Tusome. Piper, Jepkemei, and Kibukho (2015) note:

> Although the project [PRIMR] did not explicitly target the [income] poor, the basic strategies in teaching literacy and numeracy skills have proven to be effective in supporting pupils at risk for reading difficulties. PRIMR is organized in ways that align with how best to support those at risk (p. 72).

In that paper, the authors compare the positive impact of PRIMR to the negative impact of simply being poor (as measured by socioeconomic status) and conclude that the PRIMR effect is considerably larger than the poverty effect (see p. 78). This does not mean that the project definitely improved the learning of the poor, as there was no specific targeting of school support to specifically poorer regions, nor did the project work in a set of randomized poor schools and a set of randomized wealthy schools. It does mean, however, that the project's impact was enough to overcome the impact of being poor, as measured using the same dataset. At the same time, the project was able to distinguish formal from non-formal schools. The latter are more frequented by the poor, and PRIMR's impact on non-formal schools was much higher, in general, than its impact on formal schools. Effect sizes (in terms of proportions of a standard deviation) were twice as high among the non-formal schools (p. 77). But note that the effect size, in this context, is a close cousin of the coefficient of variation (the difference being that one is the inverse of the other, and in one the change in means is used as opposed to the mean itself).

We have seen above that PRIMR typically improved the coefficient of variation. This measure of "pure" inequality is thus associated with having a larger effect size among the schools more frequented by the poor.[26] Thus, the finding from Piper et al. (2015) is not necessarily inconsistent with a reduction in "pure" inequality, even if what was being reported was not the impact on pure inequality but the impact on the poor. As emphasized by a report where the impacts of various treatments are assessed, the interventions in PRIMR and Tusome were heavily focused on the basics, and also stressed fidelity of implementation (Piper, Zuilkowski, Dubeck, Jepkemei, & King, 2018). Perhaps just as importantly, both PRIMR and Tusome were fairly zealous about ensuring that the main "vectors" whereby children are helped to learn—namely the yearly scope-and-sequence of lessons, the actual lessons themselves, the books, and the formative and summative assessment—are tightly integrated with each other.[27] Indeed, Piper et al. (2018) conclude that to get the best impact you have to go "all the way", with a combination of teacher professional development, instructional support and coaching, quality student books at a 1:1 ratio, and structured, scripted lesson plans.

There are important policy, planning, and managerial implications here. Generally, if inequality is strongly driven by factors like poverty, gender, or rurality, then targeting support to schools based on those factors makes the most sense. And, after all, there are other complementary reasons to direct resources, in general, to poorer communities, as shown by the literature on income transfers. However, if there is a high degree of inequality amongst the poor themselves (and also, perhaps, inequality amongst the non-poor), then an approach that targets the basics might

[26] When the PRIMR dummy variable interacts with the poverty variable, oral reading fluency being the dependent variable, the program seems to have had greater absolute impact among the non-poor. Yet, this difference was small compared to the average (much improved) absolute level of fluency, especially in the non-formal schools. It may also be that, as noted or implied in other papers as well (e.g., Crouch & Rolleston, 2017; Crouch & Gustafsson, 2018), "pure" inequality could be reduced nonetheless, because inequality was reduced both amongst the poor and amongst the wealthier.

[27] Often, efforts to improve "quality" are more nebulous, and involve the use of "thin" inputs, such as ensuring teachers are certified, or that there is a 1:1 pupil:book ratio, without much consideration of actual teaching skill, or how relevant a book's content is.

be best—one that is integrated and executed with considerable (but not obsessive) fidelity, and (perhaps in addition) helps schools (and individual children) based on results rather than location, poverty, or gender.

Conclusions

This paper tests a variety of measures of inequality and a measure of "percent below a floor" (or, in a loose sense, "learning at the bottom of the pyramid") to see whether they are "well-behaved" with the sorts of data that are typically produced with SQC assessments. The measures, used in Kenya, include the Gini coefficient, the ratios of performance at the 90th and 75th percentiles to performance at the 10th and 25th percentiles, the coefficient of variation, the GE(2) generalized entropy measure, and the percentage of children not reading at all. These measures of inequality, and above all, changes in the measures, are then compared to average performance on the assessment and improvements in the averages.

We used the concept of "pure" inequality or dispersion to study change over time, which is assumed to be produced at least in part by random variation in teaching (where some children might be in luck and get a fairly good teacher, and others are out of luck). In some sense, this approach to inequality is one that corresponds most closely to issues such as having systems stick to standards of outcome-oriented quality assurance.

Our findings showed that the utilized measures were what we term "well-behaved". The Gini coefficient for learning, for instance, corresponded to similar numbers for income. Generally, changes (or very large or small values) in one measure correspond to changes (or the corresponding large or small values) in the other measures. Thus, there is internal coherence among all the measures and they all help to tell more or less the same story. In other words, an important conclusion from the use of the measures for assessing change over time is that the changes are strongly and consistently patterned. In addition, while it is true that, when inequality was high to start with, the reduction was greatest, it is also the case that that reduction was statistically less predictable. In the obverse, where inequality was relatively low to start with, reductions were harder to produce, but they were somewhat more certain.

In sum, we found that when one correlates improvements in the average levels (of reading fluency in this case) to changes in inequality, the larger improvements in the average almost always correspond to larger reductions in inequality. Though it is impossible to determine precisely why these inequality reductions are achieved, it seems safe to conclude that children with low initial learning (reading) results benefit disproportionately from programs that are (a) aimed at the very basics and the mechanics of learning to read; (b) contain at least the minimum necessary set of "inputs" or "vectors" of quality (e.g., teacher coaching, development of guided lesson plans, and corresponding books for children to read, at the right ratio); and (c) provide tight integration between vectors (so that lesson plans match the books' content quite rigorously, and so on for all other inputs) and are implemented with considerable fidelity. Our findings support the hypothesis that consistent measures of low-end performance can improve learning among children at the bottom of the pyramid.

References

Bruckauf, Z., & Chzhen, Y. (2016). *Education for all? Measuring inequality of educational outcomes among 15-year-olds across 39 industrialized nations* (Innocenti Working Paper 2016–08). Florence: UNICEF Office of Research.

Crouch, L., Gove, A., & Gustafsson, M. (2009). Educación y cohesión cocial. In S. Scharzman & C. Cox (Eds.), *Políticas educativas y cohesión social en Américan Latina*. Santiago de Chile: Uqbar Editores.

Crouch, L., & Gustafsson, M. (2018). *Worldwide inequality and poverty in cognitive results: Cross-sectional evidence and time-based trends* (Research on Improving Systems of Education (RISE) Programme Working Paper, RISE-WP-18/019).

Crouch, L., & Rolleston, C. (2017). *Raising the floor on learning levels: Equitable improvement starts with the tail* (An insight note from the RISE Programme). https://riseprogramme.org/sites/default/files/publications/RISE%20Equity%20Insight%20UPDATE.pdf

Dowd, A. J. (2018, March 8–13). *Visualizing learning equity: New options for communicating about learning gaps and gains* (Conference presentation). Comparative and International Education Society Annual Conference, San Francisco.

Dubeck, M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*, *40*, 315–322.

Ferreira, F. H., & Giroux, J. (2011.) *The measurement of educational inequality: Achievement and opportunity* (IZA Discussion Paper, No. 6161). http://ftp.iza.org/dp6161.pdf

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.

Graham, J., & Kelly, S. (2019). How effective are early grade reading interventions? A review of the evidence. *Educational Research Review*, *27*, 155–175.

Gromada, A., Rees, G., Chzhen, Y., Cuesta, J., & Bruckhauf, Z. (2018). *Measuring inequality in children's education in rich countries* (Innocenti Working Paper 2018–18). Florence: UNICEF Office of Research.

Hao, L., & Naiman, D. (2010). *Assessing inequality*. Thousand Oaks: Sage Publications.

Haughton, J., & Khandker, S. (2009). *Handbook on poverty and inequality*. Washington, DC: World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/11985/9780821376133.pdf

Micklewright, J., & Schnepf, S. (2006). *Inequality of learning in industrialised countries* (IZA Discussion Paper Series, No. 2517). https://www.researchgate.net/publication/5136840_Inequality_of_Learning_in_Industrialised_Countries

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. http://timssandpirls.bc.edu/timss2015/international-results/timss2015/wp-content/uploads/2016/T15-20-years-of-TIMSS.pdf

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.

Piper, B., Jepkemeib, E., & Kibukhob, K. (2015). Pro-poor primr: Improving early literacy skills for children from low income families in Kenya. *Africa Education Review*, *12*(1), 67–87. https://doi.org/10.1080/18146627.2015.1036566

Piper, B., Simmons Zuiliwski S., Dubeck, M., Jepkemei, E., & King, S. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development*, *106*, 324–336.

RTI International. (2014). *The primary math and reading* (*PRIMR*) *initiative: Endline impact evaluation: Revised edition* (United States Agency for International

Development (USAID) Working Paper Series). https://pdf.usaid.gov/pdf_docs/PA00K27S.pdf

RTI International. (2015). *Early Grade Reading Assessment (EGRA) toolkit, second edition*. Washington, DC: United States Agency for International Development.

Sahn, D. E. & Younger, S. D. (2007). *Decomposing world education inequality*. Ithaca: Cornell University. http://www.cfnpp.cornell.edu/images/wp187.pdf

Thomas, V., Wang, Y., & Fan, X. (2001). *Measuring education inequality: Gini coefficients of education* (Policy Research Working Paper, No. 2525). Washington, DC: World Bank.

Thomas, V., Wang, Y., & Fan, X. (2003). Measuring education inequality: Gini coefficients of education for 140 countries, 1960–2000. *Journal of Education Planning and Administration*, *17*(1), 5–33.

Wagner, D. A. (2003). Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research*, *39*(3), 293–309. http://authors.elsevier.com/sd/article/S088303550400031X

Wagner, D. A. (2011). *Smaller, quicker, cheaper: Improving learning indicators for developing countries*. Washington/Paris: FTI/UNESCO-IIEP. http://unesdoc.unesco.org/images/0021/002136/213663e.pdf

Wagner, D. A. (2018). *Learning as development: Rethinking international education in a changing world*. New York: Routledge.

Wagner, D., Wolf, S., & Boruch, R. (2018). *Learning at the bottom of the pyramid: Science, measurement, and policy in low-income countries*. UNESCO, IIEP.

World Bank. (2019). *Ending learning poverty: What will it take?* Washington, DC: World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/32553/142659.pdf?sequence=6&isAllowed=y