Do schools in low income countries really produce so little learning?

Sam Jones

University of Copenhagen, Denmark

20 January 2017

Abstract

The mantra ‘schooling ain’t learning’ (Pritchett, 2013) captures a view that children in many developing countries learn little in school. This study critically assesses whether such normative claims about school productivity hold-up to scrutiny. The first part sets out the challenges of evaluating the contribution of schools to achievement gains and shows existing evidence for low income countries is scant and fragile. The second part provides an original analysis of data from East Africa, covering over 700,000 children. Results from a sibling difference model reveal that the absolute contribution of schools to achievement gains is inadequate in light of national curriculum expectations. However, a new metric of relative school productivity, which evaluates achievement gains due to schooling against background rates of cognitive development associated with ageing, suggests the relative magnitude of gains due to schooling are more comparable to those in advanced countries. The gap between absolute and relative metrics of school productivity suggests that ‘learning ain’t just about schooling’. Rather, learning potential is conditioned by a wide range of background factors, including the early childhood environment.
1 Introduction

The debate around education policy in developing countries has evolved significantly over the past ten years. During the 1990s and much of the 2000s, the predominant focus was on ensuring comprehensive access to schooling. Today, while access concerns remain germane in specific cases, the quality of education is generally a bigger priority. This shift of focus toward learning outcomes is found in both academic research and policy. Goal 4 of the new Sustainable Development Goals contains an explicit commitment to ensure quality education, defined as one which leads to ‘relevant and effective learning outcomes’.\(^1\) Similarly, recent academic research highlights the importance of cognitive skills, as opposed to just access to schooling or grade progression (e.g., Hanushek and Woessmann, 2008; Woessmann, 2016).

The importance of raising the quality of education has been buttressed by claims we face a ‘global learning crisis’ (Berry et al., 2015). Data from large-scale international assessments, such as the Programme for International Student Assessment (PISA), show that average levels of educational achievement (cognitive skills) are often far lower in developing countries than in higher income countries. Surveying the range of available evidence, OECD et al. (2015) conclude that cognitive skills in “low-income countries are even farther behind high income countries than most people [have] realised” (p. 44). This claim is not disputed here. Rather, the interest of this paper is on a specific policy diagnostic that often accompanies these claims. This holds that, while many children in low and middle income countries now attend school, they learn few cognitive skills when so doing. As an illustrative example, a review of current policy debates opines that:

“... data reveals that learning levels remain critically low and that getting millions of children in school has not been translated into equal numbers of children learning the basic skills needed to thrive ... 250 million school-aged children worldwide fail to read, write, and perform basic numeracy tasks, even though half of them spent at least four years in school.” (Winthrop et al., 2015, p. 299)

The mantra ‘schooling ain’t learning’ (Pritchett, 2013) encapsulates the argument that time spent in school does not necessarily generate learning. More bluntly, the Study Group on Measuring Learning Outcomes (2013) state that “many students are learning close to nothing in school” (p.1); and OECD et al. (2015) argue that improvements in access to schooling have not led to widespread learning gains. So, a prevalent view is that an important part of the explanation for low levels of achievement is that schooling systems in many developing countries are failing to produce sufficient gains in achievement.\(^2\)

\(^1\) For the full set of SDGs and specific targets see: sustainabledevelopment.un.org (accessed January 2017).
\(^2\) This view is not merely academic but has substantially influenced donor funding to the sector. For example, the Research on Improving Systems of Education (RISE) programme, funded by the UK’s Department for International Development, is explicitly founded on a ‘growing global consensus that many developing countries face a learning crisis’. Moreover, the programme’s diagnostic point of departure is that: ‘The existing pace of improvement in
The purpose of this study is to investigate the validity of claims that school productivity in lower income countries, measured as gains in achievement due to school exposure, is indeed so inadequate. More simply, I ask: are these claims well-founded? My answer to this question is split into two parts. The first part examines existing evidence. To begin, I review the main technical issues that must be addressed to make credible normative claims about school productivity. It highlights the scope of this challenge and shows that, while accurately measuring differences in achievement levels is extremely complex, measuring the unique contribution of schools to achievement gains is even tougher and highly controversial, even in highly advanced countries. Additionally, I note that any discussion about the (in)adequacy of school productivity necessarily invokes some kind of normative benchmarks. Not only do these merit justification, but conventional benchmarks – such as comparisons against high-performing countries – are highly contested and place further demands on the comparability of achievement metrics.

The technical review, set out in Section 2, serves as a reference point from which empirical evidence about school productivity in lower income countries can be assessed. This is pursued in Section 3, which reveals that the extant evidence is rather weak. Two main criticisms stand-out. First, while few lower income countries have participated in internationally-comparable learning assessments, the learning assessments that have been carried out typically do not permit direct estimation of value-added models of school productivity, which are often preferred (Guarino et al., 2015). Second, as a consequence, many claims about school productivity are circumstantial and their precision can be questioned on methodological grounds. Indeed, much of the evidence fails to credibly isolate the unique contribution of schools to changes in achievement.

In view of these deficiencies, the second part of the paper takes a fresh look at a part of the evidence. Using five rounds of the Uwezo surveys, which comprise data on more than 700,000 school-aged children, I ask what can be said about school productivity in East Africa (Kenya, mainland Tanzania and Uganda). This is pursued in Section 4. As a first step, I explore the properties of the test score data. This reveals various concerns, such as sensitivity to aggregation choices, large temporal instability in mean test scores, and doubts regarding the comparability (equivalence) of scores over time and space. While this necessarily limits the analysis, I propose that a sibling difference model can be used to estimate broad bounds on the contribution of schools to achievement gains. On implementing this model, I find clear evidence that children are learning in school. So, the most pessimistic views about school productivity are not legitimate. However, when evaluated against the benchmark of national curriculum expectations, the pace of improvement due to school exposure (alone) does not seem adequate. Ignoring the contribution of all other inputs, it would take the average child around six years of schooling to achieve the full range of numeracy and literacy competencies expected at a grade two level.

learning is far from adequate and a decisive acceleration in the pace of improvement in learning outcomes is needed’ (RISE Vision Document 2, available at www.riseprogramme.org/sites/www.rise.ox.ac.uk/files/RISE_Vision_document-2.pdf; accessed January 2017). In turn, this view is operationalized via a focus on ‘school systems’.
The previous finding appears consistent with existing concerns about low school productivity. However, it is not self-evident that national curriculum expectations are the only appropriate yardstick. An alternative benchmark, which connects to a wider literature on interpreting effect sizes in education research (e.g., Hill et al., 2008), compares the effect of exposure to schooling to an internal ‘counterfactual’ grounded in the actual distribution of achievement. The conditional effect due to chronological ageing not only is widely known to make an important contribution to cognitive skills, but also plausibly reflects the (interactive) influence of environmental and background factors on school-readiness and learning capacities. When I compare the estimated contribution of schooling to the estimated contribution of chronological ageing, I find that relative school productivity in East Africa is commensurate with that found in advanced countries across a range of studies. These results challenge us to reflect on the extent to which schools sit in isolation from the rest of society and can compensate for the range of deprivations faced by many children in low income countries (Bernstein, 1970). The distinction between absolute and relative measures of school productivity constitutes an important contribution of this paper and further invokes the need to place school performance in context and develop empirically-reasonable expectations of their unique contribution to learning.

2 Evaluating school productivity

Turning to the measurement of school productivity, this section outlines three main issues that need to be addressed.

2.1 Measure of achievement

Concerns we face a ‘global learning crisis’ are substantiated in various ways. Among these, two main themes recur that must be viewed as distinct. The first is that levels of achievement are low on average in many developing countries. The second theme is that gains in achievement (i.e., learning), which can be attributed uniquely to school exposure, are inadequate. At the outset, therefore, any empirical assessment of either achievement or learning requires a reasonable measure of academic skills. While it is beyond the scope of the present exercise to provide a comprehensive discussion of what this entails, at least three aspects need to be considered. These are: (i) the validity of the test instrument; (ii) the reliability of the test instrument; and (iii) the properties of overall achievement metric(s) derived from individual test items. I look at each in turn.

The validity of a test instrument refers to the extent to which the items included in a test provide accurate and meaningful measures of the construct(s) of interest. A particular component of validity is the extent to which a given test adequately captures information about all relevant

3 The constructs discussed here mainly refer to basic skills in literacy and/or numeracy (see Table 1).
aspects of the chosen constructs. Reading, for instance, is widely understood to refer to a complex combination of lower and higher order capacities. These cover elements such as familiarity with letter sounds, their association to objects, prosody (reading with natural language expression), as well as various components of comprehension (e.g., vocabulary, inference etc.). Different literacy tests often focus on a sub-set of skills within the broad domain of reading (Pikulski and Chard, 2005; Keenan et al., 2008), where elements are typically chosen with reference to specific curriculum benchmarks. As such, many tests are designed to provide information about specific aspects of reading skills, ignoring either more simple or more complex tasks in that domain. As a consequence, a concrete validity challenge is to design test forms that are targetted at the right level(s). If this is not achieved then floor and/or ceiling effects can emerge, whereby children display either none or all of the competencies tested. These effects are more likely to occur where only a few test items are used to discriminate between children of different abilities and/or the same test is applied to children of highly different skill levels.

The presence of floor/ceiling effects can be particularly problematic where the aim is to accurately measure differences in achievement (over time). A stylized example illustrates. Focussing on floor effects in achievement tests, Appendix Figure A1 plots hypothetical ‘observed’ and ‘true’ achievement profiles for two countries (denoted A and B). The true profile assumes children can be differentiated along an ordinal competency scale ranging from -3 to +5. The observed scale, however, is censored from below at zero, meaning it does not discriminate between children at the low-end of the distribution. The figure shows that in grade one, the average competency of children in countries A and B are observationally similar. But this is misleading – the true average at grade one is much lower in country B, driven by a larger share of censored observations. The figure also shows the evolution of average competencies across grades. We assume all children in both countries attend school, progress normally, and gain one competency level per grade. Thus, learning gains are the same and the true profiles are parallel. In contrast, the slopes of the observed profiles diverge. This is because learning gains achieved by the (initially) lowest performing children in country B remain unobserved during the first grades. Thus, the presence of floor and ceiling effects can significantly bias measures of learning (also Resch and Isenberg, 2014).

Test instrument reliability also encompasses multiple features. A core aspect of reliability is that test results are stable across different forms (versions) of the same test, across different raters (e.g., enumerators), and across different occasions. Put more simply, if a test is reliable then test scores estimated from the same target population should not vary dramatically from one testing event to the next within a short time frame. Stability is highly desirable from the perspective of accurately quantifying trends in achievement over time. Even so, concerns around test stability are encountered among even the most technically-advanced national and international tests. In the USA, for example, Kane and Staiger (2002) show that annual changes in average scores (at the school-level) on state tests are substantial but largely driven by transient noise. As a consequence, they warn these test scores “are quite unreliable measures of performance differences across schools
and over time” (p. 99).

Temporal stability concerns also have been raised about the PISA test. Mazzeo and von Davier (2008) compare the stability of PISA results to those of the (long-term trend) National Assessment of Educational Progress (NAEP) in the USA, which is benchmark test that has been administered using the same assessment booklets, sampling, administration, scoring, analysis and quality monitoring procedures across successive assessment cycles (rounds). Their results show that absolute differences in mean scores between rounds on the NAEP are typically less that 10 percent of a standard deviation. Although variations of a similar magnitude are found on the PISA maths domain test (2003 vs. 2006), they found within-country differences in scores on the PISA reading domain between rounds were larger than any of the NAEP coefficients in around 20% of cases. The authors note that a plausible reason for this is that PISA is both a multi-lingual test and covers different content in each round. As a result, scope for variation in the relative difficulty of tests over time, especially between different language translations, appears greater (also Rutkowski et al., 2016).

Closely associated with the twin issues of validity and reliability, a third measurement challenge is how to transform a set of raw answers on a given test form to one or more quantitative metrics of achievement that, in turn, can be used to meaningfully track changes in achievement (over time). This is a perennial problem that refers to the need to identify differences in test outcomes that have a clear (equal) meaning along the full distribution of initial ability (see Braun, 1988). An ideal metric would constitute a consistent mapping of raw test results onto a unbounded scale along which unit movements have a homogeneous and precise interpretation. Unsurprisingly, this ideal is hard to attain; so, in practice, there is robust controversy regarding which methods to use.4

At risk of simplification, at least three choices must be made to arrive at a final set of achievement measures. First, is the decision regarding the number of distinct learning dimensions spanned by any given test that, in turn, merit treatment as separate outcomes. As noted above, even reading is a broad domain that comprises various competencies, such as those associated with word decoding and comprehension, which can develop at different rates. Many learning assessments span multiple general domains (e.g., reading, writing, numeracy), so the analyst must determine the number of domains to be aggregated into individual overall indicators of achievement. While this decision may be informed by ex ante test design considerations and ex post empirical measures, such as latent factor analysis, the majority of assessments tend to adopt a single overall achievement metric. However, and as Goldstein (2004) notes, this assumption of uni-dimensionality in the minimum can lead to a loss of valuable information about achievement, especially in areas that are effectively under-weighted or excluded from the final measure. Moreover, evidence suggests that violations of uni-dimensionality are not only common but also, when taken into account, can significantly alter

4 As one review study puts it: “Measures of student achievement ... are very complex devices. Although achievement is often considered as a single continuum of performance, in reality students learn many different skills and concepts and the acquisition of the skills and concepts may be at different rates. All students may not learn these skills and concepts in the same order or to the same degree of proficiency” (Reckase, 2010, p.1).
results from estimates of growth in value-added models (Lockwood et al., 2007; Goldhaber et al., 2013).

A second choice refers to the scale along which differences in achievement are to be measured. These range from simple binary proficiency scores (e.g., mastered vs. not mastered), to more complex metrics such as measures of latent ability (often denoted $\theta$) based on item response theory (IRT) models. The mapping procedures behind each of these metrics correspond to different assumptions about the data and the appropriate scale on which achievement can be expressed. Ratio or interval scales are attractive as they permit changes in achievement to be compared precisely across the full range of scores. However, valid application of such scales impose strong assumptions on the data, which may not be met in practice. This point is taken up by Ballou (2009), in reference to IRT models. He argues that available test data rarely meet the stringent conditions required to construct interval-level ability scales. Moreover, he shows that measures of changes in performance over time based on ordinal methods lead to rankings of teachers that are significantly different to those based on conventional (interval-level) measures of value added. That is, assessments of achievement and learning are sensitive to the choice of measurement scale (also Bond and Lang, 2013).

A third issue is the extent to which outcome metrics are able to account for differences in difficulty between the individual items in a test. Valuing a correct answer to a very easy question as equivalent to a correct answer to a more difficult question is intuitively inconsistent with an assumption that movements of equal distance carry the same meaning (as per an interval-level scale). Indeed, a main drawback of using ‘raw’ metrics, such as the percent correct on a test, is they do not take such differences into account. On the other hand, if we do wish to account for variation in difficulty, then some method is required to estimate item difficulty. Probability based achievement metrics, such as ridit scores (see Fielding, 1997), can be used for this purpose when test results can be placed in rank order (i.e., best to worse). IRT methods provide a more general approach to constructing scores that are sensitive to item difficulty. However, although these methods are widely employed, there is devil in the detail. Evidence suggests that analytical results, especially those that concern changes in achievement over time, are sensitive to the specific metric and/or transformation model employed (see further below; also Seltzer et al., 1994; Briggs and Weeks, 2009).

This discussion highlights that measurement of achievement is far from straightforward. Challenges become even more acute when interest focuses on differences (gains) in achievement, particularly when estimates of change are derived from tests undertaken at different points in time. Indeed, following earlier discussion, it is almost impossible to ensure that test events are perfectly equivalent in all aspects (e.g., difficulty). Moreover, it may be advisable to employ tests with different levels of content difficulty at different points in time in order to avoid floor/ceiling effects. Consequently, some form of scaling may be appropriate in order to place scores from different test forms onto a single scale from which comparisons can be consistently made. This kind of vertical scaling is a highly controversial area of applied psychometrics (Reckase, 2010) and, as always, technical
choices matter. Briggs and Weeks (2009), for instance, show that measures of growth in achievement between adjacent grades can differ by as much as 20\% of the average standard deviation across grades when different methods of vertical scaling are employed.

2.2 Model of learning

In addition to the challenges of measurement, statements about school productivity require one to isolate the unique contribution of schooling inputs to achievement outcomes. To do so, a helpful starting point is a production function for education. This posits that current achievement is a cumulative function of present and previous school, household and individual inputs (e.g., Todd and Wolpin, 2003). In general form this is given by:

\[
Y_{ihst} = \sum_{m=0}^{A_{it}} \delta_m \left( \lambda_i + \lambda_h + \lambda_t - m + C_{i,t-m} \alpha_{t-m} + X_{h,t-m} \beta_{t-m} + S_{s,t-m} \gamma_{t-m} \right) + \varepsilon_{it}
\]

which says that academic outcome (\(Y\)) of child \(i\) aged \(A\), observed at time period \(t\), who is residing in household \(h\) and attends school \(s\), is a function of: individual, family and period fixed effects (\(\lambda_i, \lambda_h, \lambda_t\)); a sequence of time-varying child characteristics (\(\bar{C}\)); a sequence of time-varying household inputs (\(\bar{X}\)); a sequence of time-varying institutional education inputs (\(\bar{S}\)), the elements of which are empty for periods when the child did not attended school; and an error term \(\varepsilon_{it}\), which aggregates all unobserved time-varying current and past factors.

Linearisation of this model yields a basic empirical specification for achievement:

\[
Y_{ihst} = \sum_{m=0}^{A_{it}} \delta_m \left( \lambda_i + \lambda_h + \lambda_t - m + C'_{i,t-m} \alpha_{t-m} + X'_{h,t-m} \beta_{t-m} + S'_{s,t-m} \gamma_{t-m} \right) + \varepsilon_{it}
\]

where \(\delta_m\) represents a time-varying discount factor that weights the contribution of inputs in different periods. Under additional assumptions (e.g., input coefficients are time invariant and the time profile of the discount factor follows a geometric progression), this specification can be further simplified to the general value-added (VA) model:

\[
Y_{ihst} = \lambda_i + \lambda_h + \lambda_t + \theta Y_{ih,t-1} + C'_{it} \alpha + X'_{ht} \beta + S'_{st} \gamma + \nu_{iht},
\]

where \(\nu_{iht} = (\varepsilon_{it} - \theta \varepsilon_{i,t-1})\)

in which \(Y_{ih,t-1}\) stands-in for past inputs; and the current discount factor, \(\delta_0\), is normalised to one. The parameter vector \(\gamma\) in equation (2) captures the (marginal) contribution of one additional unit of schooling inputs to achievement, conditional on all other current and historical inputs. While other variants on equation (1) can be contemplated, this VA specification is recognised to have
robust empirical properties and is often a preferred basis to identify the causal contribution of schooling (or other inputs) to achievement (e.g., Todd and Wolpin, 2007). As such, it highlights the extensive data requirements that may be needed to achieve credible identification, including comparable observations of achievement for the same unit at different points in time. Furthermore, not only is a complete set of current inputs required, but this VA model does not (strictly speaking) obviate the need to control for the (current) contribution of fixed factors, such as time-specific shocks and individual ability. As Guarino et al. (2015) discuss, since a complete set of control variables typically is not available to empirical researchers, estimates from VA models must be treated with ‘a high degree of caution’.

As a corollary, the framework highlights that crude estimates of differences in mean achievement over time (see Section 3) do not carry a straightforward interpretation. To see this, consider the change in average achievement for some group of children observed at time \( t \) and \( t - 1 \). Assuming the population sampled in each period is the same, this can be derived from (2) as:

\[
\bar{Y}_t - \bar{Y}_{t-1} = \Delta \bar{Y}_t = \theta \Delta \bar{Y}_{t-1} + \Delta \bar{C}_t' \alpha + \Delta \bar{X}_t' \beta + \Delta \bar{S}_t' \gamma + \Delta \bar{\nu}_t \\
= \lambda_i + \lambda_h + \lambda_t + (\theta - 1)\bar{Y}_{t-1} + \bar{C}_t' \alpha + \bar{X}_t' \beta + \bar{S}_t' \gamma + \bar{\nu}_t
\]

These expressions reveal that period comparisons of means aggregate the contributions of various determinants of achievement. If \( \theta = 0 \), which imposes that only current inputs are informative about current achievement, then (3a) implies the gain score is a weighted sum of changes in all types of inputs between periods, including the period-specific shocks and unobserved errors. Thus, for this expression to provide a reasonable approximation to the unique marginal contribution of schooling inputs, \( \Delta \bar{S}_t \) must be equal to one and all other terms roughly zero. Under the more plausible scenario where \( 0 < \theta \leq 1 \), expression (3b) suggests that the first difference of period averages will be contaminated by the lagged test score.\(^5\) The broader point is we should be sceptical of claims about school productivity (school quality) derived from estimates in which the contributions of other factors are not adequately controlled.

\[5 \text{ Note that the difference between the cases where } \theta = 0 \text{ and } \theta \leq 1 \text{ is closely related to Lord’s Paradox.} \]

2.3 Meter for comparison

A final element underlying normative claims about school productivity is a benchmark or meter against which the magnitude of specific estimates can be evaluated. Looking across existing studies, two different types of benchmarks are almost ubiquitously used. The first is direct comparisons between countries – e.g., of absolute achievement or rates of learning. Examples include PISA country rankings or bilateral comparisons of grade effects (see below). As Sellar and Lingard (2013) note, these comparisons rely on establishing reference societies against which ‘success’ is judged. A second type of benchmark is the expected level of achievement (learning progress) as set...
out in national curricula or similar reference documents.

In both of these cases, the benchmark used to evaluate achievement or learning is external in the sense it is not referenced to the *de facto* distribution of achievement (or learning) in the population to which it is applied. In itself, this is a cause for disquiet. External benchmarks are often deployed without explicit justification of *how* or *why* they constitute suitable reference points. For example, while benchmarking against other countries is often advocated for its policy impact, such comparisons naturally stimulate a focus on observable differences in policies (e.g., student tracking, teacher training, school autonomy, private provision) as an explanation for cross-country variation in learning outcomes (e.g., Woessmann, 2016). This tends to downplay the roles of more diffuse structural influences on learning, including complementaries between schooling and other background factors, which may be highly specific to each country. That is, external comparisons evoke an equivalence between countries that may be false and they risk creating de-contextualised expectations of the magnitude of feasible learning gains due to policy differences (also Alexander, 2012; Wagner et al., 2012).

A more technical concern with benchmarking against other countries is that it places additional demands on the nature of the achievement metrics. Since this form of benchmarking can only be meaningful if ‘like’ is compared with ‘like’, the cross-country equivalence of achievement or learning metrics must be secured. However, our ability to design and implement internationally-equivalent achievement tests is widely debated. Concerns around equivalence are found with respect to PISA (Grisay et al., 2007; Brown et al., 2007) but are plausibly larger across low income countries due to the great(er) linguistic and cultural diversity encompassed by these contexts. In this vein, Bartlett et al. (2015) highlights that the process and pace by which children learn to master different aspects of reading varies substantially across languages and background cultural contexts, independent of individual ability. As such, what may look like the same literacy test can elicit quite different responses and approaches in dissimilar situations. This is substantiated by Weber et al. (2015), who investigate the performance of the Peabody Picture Vocabulary Test (PPVT) among young children in Madagascar. They find significant variation in item functioning across the various dialects of Malagasy, as well as in accordance with the child’s daily exposure to French. As such, the authors conclude that: “the success of optimizing an instrument for a given context will depend on being able to change, drop, add, and re-arrange items as necessary” (p. 17). A similar reservation is noted by Cueto et al. (2012) in their review of the battery of tests developed for the third round of the Young Lives surveys. Their analysis of the properties of the test items lead them to conclude that: “the process of adapting tests did not perform a sufficiently thorough process to ensure that the levels of difficulty are comparable across languages” (p. 34).

The challenge of establishing equivalence indicates a critical tension between the local and inter-

---

6 This is substantiated, informally, from the range of vocabulary included in the Indian Young Lives PPVT test. This included terms such as: ‘raccoon’, ‘penguin’, ‘cimcorder’ and ‘knight’, all of which may be appropriate in a relatively affluent Western context but appear less suitable for the target population of poorer children in India.
national validity of learning assessments (Wagner, 2010). Tests that seek to ensure international validity may encounter bias due to variation in item functioning and weak targeting of test instruments to local skill levels. However, efforts to avoid these problems and enhance local validity risk undermining the direct comparability of outcome metrics across contexts. In this light, benchmarking against standards set out in national curricula would seem to be an attractive option – i.e., avoiding the need for international equivalence. Even so, a running theme of scholarship critiques curricula in many developing countries as being detached from local conditions, over-ambitious, and inimical to achieving rapid learning gains. This critique has gained force in light of very large increases in access to schooling over recent decades. That is, as (primary) schooling is no longer the exclusive domain of relatively wealthy households, so learning expectations may need to be adapted to account for widely different levels of school-readiness and family academic support capacities. In line with this view, Pritchett and Beatty (2015) suggest low achievement gains in schools are driven by an excessively fast expected pace of the curriculum. Evidence from teachers in various developing countries further confirms that curriculum expectations are often over-ambitious and fail to give due weight to foundational competencies (Altinyelken, 2010; Waller and Maxwell, 2016; Erling et al., 2016). As such, it is not self-evident that learning gains should be exclusively benchmarked against national curriculum expectations.

3 Existing evidence on learning

With the challenges of the previous section in mind, I now appraise the range of empirical evidence that has been used to support claims that school productivity is inadequate in many low income countries. To begin, much of the debate about international differences in school quality relies on data from large-scale international learning assessments (ILAs). These tests, covering both primary and secondary school students, have grown from coverage of just a handful of high income countries in the 1970s to a wide range of countries today. This expansion is illustrated in Figure 1, which shows the number of countries participating in ILAs over the period 1965-2010. Despite rapid growth in participation, however, coverage of lower income countries remains thin. For example, only four low income countries have collaborated in the PISA; and only three sub-Saharan African countries have ever participated in the Trends in International Mathematics and Science Study (TIMSS), all of which are now classified as middle income nations (Botswana, Ghana and South Africa).

Putting to one side multi-region (global) international assessments such as PISA and TIMSS, a range of more focussed regional tests have been deployed extensively in (poorer) developing countries, especially in recent years. These assessments are summarised in Table 1, from which three points merit comment. First, these kinds of assessments have been largely targeted at children of primary school age. This reflects the reality that enrolment in secondary school is more limited in lower income contexts. It also corresponds to concerns that fundamental skills in reading and
Numeracy are not adequately mastered by children attending primary schools.

Second, the table suggests that not all assessments share the same objectives. The first four assessments incorporate an explicit commitment to provide internationally- and temporally-comparable measures of student achievement. Consequently, substantial effort is invested to standardize the test instruments, meaning the test forms deployed in different countries and at different times incorporate many of the same questions, allowing the forms to be formally equated. This objective is similar to that of other ILAs, such as PISA, and require significant technical capacity to implement. The other assessments in the table, however, are not primarily oriented towards providing internationally-comparable metrics of achievement. For instance, the Early Grade Reading Assessment (EGRA) is principally a diagnostic tool to evaluate foundational skills in reading (Dubeck and Gove, 2015). Indeed, while the application of the tool in different contexts (languages) adheres to a common framework, the tests are typically adapted to fit local circumstances (see also LMTF, 2013). The same goes for the Citizen-Led Assessments (CLAs; see further below; also Tobin et al., 2015), which are designed according to a common template but are tailored to suit national curricula. Consequently, as the table notes, these assessments can be viewed as being only partially standardized, meaning they are loosely but not strictly comparable between countries.

Third, reflecting differences in aims, the target populations vary considerably across the various
Table 1: Summary of international learning assessments deployed in lower income countries

<table>
<thead>
<tr>
<th>Type/Name</th>
<th>Domains¹</th>
<th>Std.ized?</th>
<th>Objective</th>
<th>Target population</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIRLS</td>
<td>R</td>
<td>Yes</td>
<td>Comparative</td>
<td>4th-6th grade</td>
<td>51</td>
</tr>
<tr>
<td>PASEC</td>
<td>R, N</td>
<td>Yes</td>
<td>Comparative</td>
<td>2nd &amp; 5th grade</td>
<td>13</td>
</tr>
<tr>
<td>SACMEQ</td>
<td>R, N</td>
<td>Yes</td>
<td>Comparative</td>
<td>6th grade</td>
<td>15</td>
</tr>
<tr>
<td>LLECE</td>
<td>R, N, W, S</td>
<td>Yes</td>
<td>Comparative</td>
<td>3rd &amp; 6th grade</td>
<td>15</td>
</tr>
<tr>
<td>EGRA</td>
<td>R</td>
<td>No</td>
<td>Diagnostic</td>
<td>1st-4th grade</td>
<td>&gt;70</td>
</tr>
<tr>
<td>EGMA</td>
<td>N</td>
<td>No</td>
<td>Diagnostic</td>
<td>1st-3rd grade</td>
<td>22</td>
</tr>
<tr>
<td>CLAs</td>
<td>R, N</td>
<td>No</td>
<td>Diagnostic</td>
<td>School-aged children</td>
<td>13</td>
</tr>
<tr>
<td>Young Lives</td>
<td>R, N, V, W, I</td>
<td>Yes</td>
<td>Longitudinal</td>
<td>Various ages</td>
<td>4</td>
</tr>
</tbody>
</table>


Note: ‘Std.ized’ indicates whether identical tests (or translated equivalents) are administered across different countries in each round; ‘N’ gives the approximate number of countries covered by each assessment to date.

assessments. In fact, only one of the assessments described in Table 1 can be used to track achievement at the individual-level over time. With the exception of the Young Lives studies, all other assessments are cross-sectional in nature, meaning that (under standard designs) there is no attempt to assess the same children or schools across different periods. Additionally, assessments that focus on achievement in specific grades typically do not even allow changes in average outcomes for the same target group of pupils to be monitored. Thus, referring back to the data requirements implicit in equation (2), the design of many assessments for which we actually have data do not facilitate any straightforward implementation of models to credibly measure school productivity.⁷

The majority of evidence used to argue that school productivity is inadequate in low income countries falls into two categories. The first relies only on estimates about levels of achievement at single points in time. As indicated in Section 1, a wide range of ILAs suggest that average achievement in many lower income countries is below that of developed countries (or specific top-performing nations such as Vietnam). Gove and Cvelich (2011), for instance, summarise results from various EGRAs and report that a majority of children finishing grade two could not read a single word in a relevant language of instruction in numerous countries (e.g., Mali, Uganda, Gambia). Reports from the ASER assessments in India similarly reveal that only 40 percent of children attending grade three in 2014 were able to read a grade one level text and only 25 percent were able to perform basic subtraction (ASER Centre, 2015). Data from East Africa, reviewed in Section 4, tell a similar story; and, in a recent exercise, Sandefur (2016) uses a small overlap between the SACMEQ and TIMSS tests to equate their scores onto a common scale. He finds that students from many SSA countries score more than 100 points (one standard deviation) below students from high income countries (also Barro and Lee, 2015; OECD et al., 2015).

Substantial international differences in absolute achievement levels provide a logical basis to argue

⁷ PISA data, for instance, is not generally used to isolate the unique contribution of schools to achievement. By way of exception, Bratti and Checchi (2013) retest Italian participants in PISA in order to estimate VA models.
that school productivity is inadequate. That is, assuming this evidence is correct, then children in poorly-performing contexts do need to learn ‘more’ to close achievement gaps (see Study Group on Measuring Learning Outcomes, 2013); and schools are a natural focus for where this extra learning can be produced. Similar arguments are made using the gap between observed achievement and the expected academic standards for the grade in which the child is enrolled. Where surveys show many children are performing below grade-expectations, as stipulated in national curricula, a corresponding diagnosis is that current rates of progression in school are inadequate (e.g., Pritchett and Beatty, 2015; Banerjee et al., 2016). A more elaborate but analogous argument is made by Beatty and Pritchett (2012). They use changes in mean achievement for a given target population viewed over separate assessment rounds (e.g., PISA) to estimate improvements in school effectiveness over time. They highlight that, at current rates of improvement, many developing countries would take generations to close the international achievement gap.

These gap-type arguments raise important issues. But, considered as normative claims about school productivity, they can be critiqued on three grounds. First, without imposing additional assumptions, absolute gaps say nothing about rates of achievement gains over time. It could be the case that children in lower income countries progress in school at roughly the same pace as children from richer countries in a *ceteris paribus* sense, the only difference being that the former start from a lower baseline or receive fewer complementary inputs outside school. Indeed, Glewwe et al. (2001) find the productivity of a year of schooling is positively influenced by enhanced early childhood nutrition. If this is so, then while it may still be accurate to claim such children need to learn ‘more’ to close international achievement gaps, this does not necessarily imply that schools alone are producing inadequate gains in achievement.

A second and related critique corresponds to the discussion of Section 2.3. Gap-type arguments do not isolate the causal contribution of schooling to learning. Where they are used to make inferences about school productivity, an implicit assumption must be that schools are a dominant and meaningfully separable determinant of differences in achievement. Although not implausible, such a view is controverted by further evidence that non-school factors, such as the home environment, play a critical role in determining the academic performance of young children (Heckman, 2006; Björklund and Salvanes, 2011; Jones and Schipper, 2015). Minimally, implicit auxiliary assumptions need to be explicitly justified before policy inferences are drawn.

The third critique of gap-type arguments relates to the suitability of applying benchmarks from high-performing (advanced) nations. These concerns were also raised in Section 2.3 in general terms. A specific point here is that there is little direct overlap between the test instruments (ILAs) deployed in low income countries and those in developed countries – i.e., achievement measures

---

8 These arguments are not just made in relation to low income countries. There is a long tradition of research in the USA that raises concerns about low performance at the state- and national-levels. For example, a statement to the : “In the State of California, 59% of fourth grade children had little or no mastery of the knowledge and skills necessary to perform reading activities at the fourth grade level, compared to a national average of 44% below basic reading levels. ... reading failure is a serious national problem.”
are not strictly equivalent. Furthermore, since most ILAs available for the former set of countries are not designed to support precise international comparisons, there is likely to be substantial uncertainty regarding the precise magnitudes of international achievement gaps even among lower income countries.

The second category of evidence that makes the case school productivity is inadequate uses differences in average achievement on the same or similar tests administered to children enrolled in different grades (often at one point in time). Pritchett and Beatty (2015) use results from a range of CLAs to remark that only a small share of children master new skills as they proceed to higher grades. Their argument is that in many lower income countries, children display shallow learning profiles, defined as the slope of achievement gains across grades, which in turn substantiates evidence that a small share of children meet academic grade-expectations (Pritchett, 2013). Estimates of this sort are certainly informative, but many of the earlier concerns apply. In particular, as shown in Section 2.2, direct comparisons of means are unlikely to provide a reasonable approximation to the unique contribution of schools to differences in achievement. Similarly, Rolleston and James (2015) use the Young Lives data to compare average maths performance in India and Vietnam for children observed in 2006 and in 2009. Whilst this has the advantage of using data for the same individuals, the authors recognise that the underlying maths tests are not directly comparable over time. Thus, their estimates cannot be used to make (comparative) claims about rates of progression.

In sum, much of the evidence about school productivity in low income countries is circumstantial in nature. Rather than relying on rigorous causal estimates of the unique contribution of schooling to achievement, inferences about learning are often drawn from evidence on absolute levels or mean differences in achievement. Thus, combined with the inherent technical difficulties involved in establishing directly-comparable achievement metrics, these inferences are open to question. A response to this position comes from the limited number of studies that do employ methods to control for the contribution of other factors. Rolleston (2014), for instance, exploits the longitudinal aspect of the Young Lives assessments to compare the contribution of differences in years of schooling to achievement, conditional on other variables. He finds large differences between the four countries, particularly lower productivity in Ethiopia and higher productivity in Vietnam. Singh (2016) uses similar data and makes further efforts to equate test scores between countries at each point in time. He also finds that schooling is significantly more productive in Vietnam than in Peru. However, since the underlying tests used in these studies are not equated within each country onto a consistent scale, the requirements of the VA model given in equation (2) are not fully met.9 Additionally, a drawback of the Young Lives assessments is that the samples were taken from purposively-selected locations, making both within-country generalization and cross-country comparisons problematic. Hence, the degree of uncertainty in the estimates associated with sampling and measurement issues is unclear and, again, cautions against drawing strong

9 Specifically, as a measure for early achievement, Singh (2016) uses a general cognitive development score; but for later achievement (the dependent variable) he uses a separate maths test score (with no linking items). As such, a maintained assumption must be that both outcomes conform to an identical production function.
conclusions.

4 New evidence for East Africa

The previous section raised questions regarding the adequacy of evidence about school productivity in low income countries. Perhaps surprisingly, it noted that few detailed studies have been undertaken to rigorously and directly estimate the causal contribution of schooling to achievement, particularly using the kind of nationally-representative data that has been collected by large-scale assessments in low income countries (see Table 1). The remainder of this paper takes steps to address this gap, focusing on three East African countries where the Uwezo initiative has collected learning assessment data annually since 2010. The main question of interest is: what can be said – both positively and normatively – about rates of learning in these countries?

Following the review of Sections 2 and 3, any answer to this question must begin with an examination of the properties of the Uwezo test data. This informs the choice of estimation strategy and indicates potential limitations as regards the expected accuracy and robustness of the results. Thus, after briefly describing the Uwezo surveys, Section 4.1 examines specific aspects of validity and reliability. Section 4.2 goes on to exposit and implement a sibling fixed effects model, which is used to approximate the causal contribution of schooling to learning. Finally, 4.3 normatively evaluates the magnitude of the resulting estimates using various yardsticks, including what I call an internal benchmark, which is the counterfactual rate of learning associated with chronological ageing and from which a metric of relative school productivity is derived.

The Uwezo data refers to the set of large-scale household surveys undertaken in Kenya, mainland Tanzania and Uganda conducted by the Uwezo initiative annually since 2010 (for further details see Uwezo, 2012; Jones et al., 2014). The approach adopted by Uwezo was inspired by exercises carried out in India by the Assessment Survey Evaluation Research Centre (ASER). Since 2005, ASER has used a network of volunteers to survey the literacy and numeracy abilities of over 500,000 children each year. As with ASER, the Uwezo surveys are citizen-led assessments and have been designed following a consistent format. The target population has been children of school-age residing in households (not institutions), up to age 16. The surveys are representative at the national and district levels, based on the administrative classifications in the most recently available population census. In some survey rounds, however, administrative difficulties meant that certain districts could not be surveyed. Throughout, (adjusted) survey weights are used that take into account these implementation issues.

In each assessment, information was collected at the household-level, covering household characteristics and details of all children (e.g., age, gender, whether or not attending school etc.). Within each surveyed household, children aged 6-16 also were individually administered a set of basic...
oral literacy and numeracy tests. These tests have been based on a common template, following the original ASER model, but have been tailored to each country and varied by survey round (so as to avoid learning effects). Specifically, in each round and country, local experts have taken the template and developed item content to reflect competencies stipulated in the national curriculum at the grade 2 level. That is, the tests are criterion-referenced to skills that should be achieved by the majority of pupils after two years of completed schooling.

The literacy tests refer to the languages in which pupils are tested at the end of primary school – namely, English and Kiswahili in Kenya and Tanzania; and only English in Uganda.11 These languages are not necessarily the same as the (primary) language used for teaching since various language-of-instruction practices are found across the region. The Uwezo literacy tests evaluated simple reading skills in order of increasing difficulty. Based on pre-prepared test cards, children were asked to: recognise a letter from the alphabet, read a word, read a sentence, and read a paragraph (story). Provided the child was able to read at the story level, she was further asked at least one question to assess whether she also comprehended the story. Thus, based on this collection of items, the child is scored on an ordinal scale ranging from 0 to 5. The numeracy skills covered number recognition, counting, and the performance of basic calculations with numbers of up to two digits (addition, subtraction and multiplication). Children were asked a set of questions (also from pre-prepared cards) starting with simple arithmetic and either increasing in difficulty, if they were successful, or decreasing in difficulty if not. As with the literacy tests, the design of the numeracy tests means they correspond to a raw score on an ordinal scale.

4.1 Validity and reliability

Following Section 2, various properties of the data merit investigation. To begin, I look at construct validity – namely, the potential for bias if test instruments are not well targeted to respondents’ skill levels, leading to significant floor or ceiling effects. The basic level at which the tests are set immediately raises the prospect of ceiling effects. So, I start by excluding all older children from the sample. Indeed, the analytical sample used hereafter only includes children who are age-appropriate to be attending grades 1 to 5 in their respective country – i.e., 6-11 year olds in Kenya and Uganda; and 7-12 year olds in Tanzania. Based on this sample, Table 2 reports the share of children classified at either the floor or ceiling of the raw ordinal scale in each skill domain of the Uwezo tests. The top half of the table uses the full sample of children (i.e., all children expected to be attending grades 1–5 on an age basis). The bottom half of the table restricts the sample to children attending grade 3, which is the group for whom the tests should be most suited.

Even after excluding older children, the table reveals that floor and ceiling effects are both prevalent. For the full sample, more than one in three children is classified at an extreme value in at least

---

11 Examples of the tests used in the survey are available from www.uwezo.net.
Table 2: The prevalence of floor and ceiling effects in raw scores (% sample)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Domain</th>
<th>Level</th>
<th>KE</th>
<th>TZ</th>
<th>UG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>English</td>
<td>Floor</td>
<td>10.4</td>
<td>45.7</td>
<td>35.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>23.7</td>
<td>9.3</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>Math</td>
<td>Floor</td>
<td>8.9</td>
<td>18.7</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>24.9</td>
<td>24.6</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>Swahili</td>
<td>Floor</td>
<td>14.1</td>
<td>26.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>26.0</td>
<td>25.5</td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>English</td>
<td>Floor</td>
<td>2.5</td>
<td>33.9</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>31.4</td>
<td>11.6</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>Math</td>
<td>Floor</td>
<td>1.7</td>
<td>7.6</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>37.2</td>
<td>35.6</td>
<td>23.2</td>
</tr>
<tr>
<td></td>
<td>Swahili</td>
<td>Floor</td>
<td>4.3</td>
<td>12.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ceiling</td>
<td>35.3</td>
<td>36.5</td>
<td></td>
</tr>
</tbody>
</table>

† only includes children currently enrolled in grade 3.

Note: full sample includes 6-11 year olds in Kenya and Uganda and 7-12 year olds in Tanzania; cells report the share of children at highest (ceiling) or lowest (floor) levels of attainment; KE is Kenya; TZ is Tanzania (mainland); and UG is Uganda.

Source: own estimates.

one individual domain. In Tanzania, the majority of children are located at the extremes in the Swahili test. Floor and ceiling effects remain prevalent among children attending grade 3. It follows that estimates of learning gains may be biased (downwards) but also that the extent of this bias is unlikely to be consistent across countries. For example, for the full sample, floor effects are largest in Uganda; but, ceiling effects are greatest in Kenya. Despite this, the composite score appears somewhat less affected by these dynamics, precisely because it encompasses more information about each child.

A second property of the data is the stability (reliability) of test scores over time. It was already noted that many large-scale assessments display non-negligible variations in average scores across consecutive test rounds. The Uwezo assessments are no different. The top row of Figure 2 illustrates the means for each country and survey round, transformed into percent correct scores and, again, restricting the sample only to children attending grade 3. The results clearly show there has been material variation in mean scores across rounds within each country. For example, in Uganda in 2013 the average child attending grade 3 scored 59% on the numeracy test, which was about 12 percentage points higher than in the previous year. Indeed, across most countries and tests, the highest and lowest means over the five rounds cover a spread of around 10 percentage points. Consequently, between-country gaps in mean scores vary substantially in magnitude (and sometimes switch sign) depending on the test rounds chosen for comparison.

Variation in mean scores across rounds is encountered if one uses other scoring functions. One of these is the ridit, which uses the empirical distribution of ordered raw scores to place results on a
probability scale. Specifically, the ridit score for category $j$ of an ordinal scale is given by:

$$r_j = 0.5p_j + \sum_{k<j} p_k$$ (4)

where $p_j$ is the probability associated with observing a response in category $j$. Often likened to percentiles, Fielding (1997) demonstrates ridits represent conditional mean scores (i.e., conditional on the ordinal level achieved by the child), based on an underlying uniform distribution constrained between zero and one. As such, ridits proxy for the cumulative probabilities that would be observed if achievement followed a continuous scale. Calculating the ridits separately for each country-round combination, the bottom row of Figure 2 plots the mean ridit for children attending grade 3 (only).

By construction, the full sample means are constrained to equal 50 ridits in each country and round. Since this imposes a kind of equivalence across tests, it is unsurprising the grade 3 mean ridits show less variation relative to the percent correct scores. Nonetheless, the means do vary in a material fashion – e.g., in 2015, the average Tanzanian child attending grade 3 scored 67 ridits in the numeracy test, versus 59 in 2013. Since the standard error of the numeracy ridit mean in Tanzania is consistently less than one point, this kind of variation appears highly significant.
Table 3: Indicators of test score reliability within grades between years, by district

<table>
<thead>
<tr>
<th>Metric / Variable</th>
<th>Variation in changes</th>
<th>% Transitory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KE</td>
<td>TZ</td>
</tr>
<tr>
<td>Numeracy (%)</td>
<td>13.2</td>
<td>18.9</td>
</tr>
<tr>
<td>English (%)</td>
<td>15.4</td>
<td>36.6</td>
</tr>
<tr>
<td>Swahili (%)</td>
<td>15.6</td>
<td>24.9</td>
</tr>
<tr>
<td>Numeracy (ridit)</td>
<td>12.5</td>
<td>13.7</td>
</tr>
<tr>
<td>English (ridit)</td>
<td>12.1</td>
<td>13.6</td>
</tr>
<tr>
<td>Swahili (ridit)</td>
<td>12.4</td>
<td>12.8</td>
</tr>
<tr>
<td>Age</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Female</td>
<td>16.4</td>
<td>16.0</td>
</tr>
<tr>
<td>No. children in hhld</td>
<td>11.4</td>
<td>15.5</td>
</tr>
<tr>
<td>Household is non-poor</td>
<td>19.9</td>
<td>24.7</td>
</tr>
</tbody>
</table>

Note: estimates are based on grade-specific means for each district in each survey round; only districts covered in all rounds are included in the estimation sample; children not attending school are excluded; KE is Kenya; TZ is Tanzania (mainland); and UG is Uganda.
Source: own estimates.

Plausibly, the variation in means scores across rounds may be driven by multiple factors. These include systematic changes in performance (i.e., genuine change), sampling error, and various forms of non-sampling error such as that induced by differences in the underlying difficulty of the test forms. To explore this further, Table 3 reports a metric of the variation in the first difference (annual change) in mean scores within each grade, calculated at the district level. Specifically, for each country the first column reports the standard deviation of the first difference normalised by the pooled mean level (covering all rounds). The second column of the table follows Kane and Staiger (2002) and estimates the proportion of variation in annual changes (again, within each grade) that is due to transitory noise. This is estimated as minus two times the correlation between the changes in scores in one year and changes in scores the next year. The first part of the table reports these metrics for the percent correct test scores, the second part for the ridit scores and, for comparison, the third part calculates them for a number of background variables.

Three findings stand out. First, consistent with the insights of Figure 2, the normalized variation in test scores appears substantial. For all test metrics, the standard deviation is more than 10% of the mean score level and in some cases is more than 20%. By comparison, the variation of the average age for children in each grade is lower by a factor of almost 10. Even so, there seems to be comparable levels of variation in other background variables including the gender composition of grade cohorts and estimates poverty levels. Consequently, we cannot discount that either random or systematic differences in the samples across rounds explains part of the variation in test scores. The third finding is that the vast majority of between-round variation appears to be driven by transitory noise rather than genuine changes. For the composite test score metrics, for example, at least 60% of the annual variation in means is not permanent. These magnitudes are similar for background variables, which also are not expected to be trending systematically over the period analysed.
A final property of the data is the comparability of the test scores over time and space. In order to meaningfully compare achievement (levels and gains) on an international basis, a necessary assumption is that the test forms used in different countries measure the same ‘thing’ – i.e., children of the same underlying ability should attain approximately the same score, regardless of the specific test form deployed. While the Uwezo tests are not designed to be internationally-equivalent, they follow a common template and structure. As such, and especially since they cover similar basic skills (see above), they may be comparable. Item response theory (IRT) provides a formal window on this property via analysis of differential item functioning (DIF). Analysis of DIF is typically used during assessment development and pre-analysis phases to identify specific items that appear biased and which may contaminate estimated final scores. An underlying assumption of IRT methods is that item parameters are independent of the particular sample from which they are estimated. Thus, as long as an item is functioning in the same way (i.e., is valid), corresponding parameter estimates should be stable across different contexts/samples.

To proceed, I take data from the 2013 assessment rounds and estimate graded response models on the English literacy and numeracy items (found in all assessments but with differing content). To focus purely on item functioning, I restrict the sample to the core target group of children attending grade 3. With this sample, I estimate a restricted graded response model, assuming the difficulty and discrimination parameters of the model are the same for all children in each country – i.e., I run a pooled model. Next, I run an unrestricted model, allowing the parameters to vary by country. The results from this exercise are shown in Appendix Figure A2. The figures are plot boundary characteristic curves (BCCs), which show the probability of correctly answering the item at a specific level, conditional on the estimated ability of the respondent (denoted theta). Consistent with the structure of the graded response model, each curve represents a single level or category; and the levels are placed in order of increasing difficulty as one moves from left to right (or as theta increases). The vertical dashed lines indicate the value of theta that corresponds to a 50% probability of attaining a specific level. The BCCs point to substantive differences in item functioning across the countries.

Taking the English reading test to illustrate (Figure A2a), the value of theta associated with a 50% probability of correctly answering at the highest level is close to three in Uganda but less than two in Kenya. Moreover, the gap between the highest and lowest thetas associated with the 50% probability cut-offs vary substantially – being more than 6 in Kenya but around 4 in Tanzania. These results cast doubt on the suitability of making direct international comparisons of either achievement or learning based on the raw test scores or monotonic transformations thereof. This is confirmed by likelihood ratio tests, which formally assess whether the constraints implicit in the restricted model are valid. The null hypothesis that the homogeneity restrictions are valid is rejected for both skill domains.

A final insight into comparability is given by an analysis of DIF within countries over time – i.e., the interest here is whether the different test forms employed in consecutive survey rounds function
in a consistent way. Results from this analysis, which follows the approach outlined above, are
given in Appendix Figures A3–A5. Formal likelihood ratio tests reject the assumption of parameter
homogeneity in all cases (tests and countries). Visual analysis of the BCCs confirms there are stark
differences in item difficulty over the rounds. An obvious case is numeracy in Uganda – in 2012, a
child attending grade 3 with an ability value of around 3 was predicted to pass the highest level
of the test with 50% probability; but in 2015, a child with an ability value of less than two was
predicted to do the same.

4.2 Empirical strategy

The previous sub-section raised concerns about the validity and reliability of the Uwezo tests. At a
minimum, it implies caution must be exercised when making direct comparisons of scores over
time – e.g., gains in percent correct scores are unlikely to provide accurate measures of genuine
changes in (mean) achievement. Two implications follow. First, it may be appropriate to focus on
scores that are scaled in a consistent fashion across countries and rounds so as to reflect apparent
differences in form functioning (item difficulty). However, since there are no anchor items across
the tests (i.e., no identical questions) and no homogeneous (retested) sub-groups who might also
serve to anchor scores, I rely on ridit and estimated IRT scores calculated separately for each
country and round (see further below). For these metrics, scores will not be directly comparable
over time or space; but variation between pupils within a given country and round can be interpreted
on a consistent basis. That is, these scores are always meaningful relative to the distribution of
achievement observed for a specific test event. In turn, second, it is appropriate to focus analysis
on estimating school productivity based on data for specific countries and survey rounds (i.e., in
cross-section), rather than via temporal variation.

Taking on board these suggestions, the analysis proceeds by employing a range of achievement
indicators. These differ according to their measurement scales, as well as the extent to which they
capture variation in item difficulty. The first metric is the percent correct score, which assumes the
raw ordinal scores from the tests can be treated as cardinal and form-invariant (i.e., a child scoring
2 on a Kenyan test would score a 2 on a Tanzanian test). The second metric is the ridit score and the
third is the IRT ability score, both described in Section 4.1. The latter two metrics are calculated
both on a pooled and a separate basis (individually for each country and survey round). These
conform to diametrically opposite assumptions about test form equivalence; and, as such, provide a
valuable basis to compare whether analytical results are sensitive to such alternative assumptions.

To ensure consistent treatment of the various competencies spanned by the Uwezo tests, the above
scoring functions are applied separately in each skill domain. Since the IRT scores have no
natural scale, they are standardized to take a mean of fifty and standard deviation of 100 for the

12 Further detail about the construction of the metrics discussed in this sub-section is available on request from the
author.

21
Table 4: Alternative composite metrics of achievement

<table>
<thead>
<tr>
<th>Metric →</th>
<th>% Correct</th>
<th>Ridit</th>
<th>IRT</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Pooled</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>KE</td>
<td>52.7</td>
<td>59.4</td>
<td>50.0</td>
<td>83.3</td>
</tr>
<tr>
<td></td>
<td>(32.2)</td>
<td>(24.1)</td>
<td>(26.4)</td>
<td>(85.8)</td>
</tr>
<tr>
<td>TZ</td>
<td>40.9</td>
<td>47.2</td>
<td>50.0</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>(31.1)</td>
<td>(25.2)</td>
<td>(24.3)</td>
<td>(89.3)</td>
</tr>
<tr>
<td>UG</td>
<td>31.8</td>
<td>42.7</td>
<td>50.0</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td>(28.0)</td>
<td>(23.4)</td>
<td>(25.5)</td>
<td>(82.5)</td>
</tr>
<tr>
<td>All</td>
<td>41.1</td>
<td>49.1</td>
<td>50.0</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>(31.5)</td>
<td>(25.2)</td>
<td>(25.4)</td>
<td>(89.2)</td>
</tr>
<tr>
<td>KE-UG</td>
<td>69.1</td>
<td>70.2</td>
<td>0.0</td>
<td>69.9</td>
</tr>
<tr>
<td>KE-TZ</td>
<td>37.3</td>
<td>49.5</td>
<td>0.0</td>
<td>50.0</td>
</tr>
<tr>
<td>TZ-UG</td>
<td>30.6</td>
<td>18.3</td>
<td>0.0</td>
<td>17.5</td>
</tr>
</tbody>
</table>

Note: Cells show means and standard deviations (in parentheses); final column is the number of observations included in the analytical sample; KE is Kenya; TZ is Tanzania (mainland); and UG is Uganda; pooled metrics are calculated using all rounds and countries simultaneously; non-pooled metrics use each round separately in each country; IRT scores are standardized to the distribution(s) of all children in the relevant sample; final rows give bilateral country gaps in effect size units. Source: own estimates.

observations used in their construction. Thus, for the non-pooled scores, the mean in each country and survey round is the same for both the ridit and IRT scores. As a final step, I then aggregate scores across competencies to yield a composite measure of achievement. As before, this is given by the row-wise average of all available competencies and is constructed in the same way for each of the five aforementioned metrics (1 × percent correct + 2 × ridit + 2 × IRT scores). Although simple, this approach is supported by factor analysis. For each country, the first principal component of the raw test scores accounts for around 80% of the variation in the sample and suggests approximately equal weights can be given to each competency (domain).

Table 4 summarises the means and standard deviations of the five composite metrics. Notably, the pooled metrics all indicate that mean achievement in Kenya appears to be consistently superior to that of the other countries, while mean achievement in Uganda appears always inferior. That said, these metrics do not yield identical insights. The magnitude of achievement gaps between countries, when considered in effect size units, is sensitive to the choice of scaling function. For instance, the percent correct score suggests the mean achievement gap between Tanzania and Uganda is substantially larger than under the ridit and IRT scores. This confirms that the choice of metric matters.

An outstanding challenge is how to estimate school productivity. While Section 2.2 noted that VA estimates are often preferred, most data from large-scale learning assessments, including that

\[ \sqrt{2(\mu_i - \mu_j)/\sqrt{(\sigma_i^2 + \sigma_j^2)}}, \]

where \( \mu_i \) is the test score mean for country \( i \) and \( \sigma_i \) is the standard deviation.

13 Following Briggs and Weeks (2009), the gap in effect size units is: \( \sqrt{2(\mu_i - \mu_j)/\sqrt{(\sigma_i^2 + \sigma_j^2)}} \).
collected from the Uwezo surveys, is not longitudinal in nature. Moreover, the presence of a lagged test score in VA models may be particularly problematic in the present setting due to large (noisy) temporal variation. As an alternative, Todd and Wolpin (2003) show that models using observations of (biological) siblings at the same point in time can also be used to estimate the parameters of education production functions. This approach is feasible here since the Uwezo data was collected at the household-level and includes test score data for all school-age children resident in the same household. To see how the sibling estimator works, consider the first difference of equation (1) taken over two sibs ($i = 1, 2$) aged $A$ and $A - 1$:

$$Y_{1\text{hst}} - Y_{2\text{hst}} = \delta_A \left( \lambda_i + \lambda_h + \lambda_{t-A} + C'_{1,t-A}\alpha + X'_{h,t-A}\beta + S'_{s,t-A}\gamma \right)$$

$$+ (\lambda_1 - \lambda_2) \sum_{m=0}^{A-1} \delta_m + \beta \sum_{m=0}^{A-1} \delta_m (X_{1,t-m} - X_{2,t-m})'$$

$$+ \alpha \sum_{m=0}^{A-1} \delta_m (C_{1,t-m} - C_{2,t-m})' + \gamma \sum_{m=0}^{A-1} \delta_m (S_{1,t-m} - S_{2,t-m})'$$

$$+ (\varepsilon_1 - \varepsilon_2)$$

Next, consider the restrictions: $\delta_A \approx 0$, which says that temporally distant inputs have a negligible effect on current achievement; and $\forall m: (X_{1,t-m} - X_{2,t-m} \approx 0, C_{1,t-m} - C_{2,t-m} \approx \bar{C}_\Delta, S_{1,t-m} - S_{2,t-m} \approx \bar{S}_\Delta)$, which says that household inputs are the same for all siblings (when both alive), while the difference in child and school inputs is approximately constant. Under these restrictions, the previous expression simplifies to:

$$Y_{1\text{hst}} - Y_{2\text{hst}} \approx (\lambda_1 - \lambda_2) \sum_{m=0}^{A-1} \delta_m + \alpha \bar{C}_\Delta \sum_{m=0}^{A-1} \delta_m + \gamma \bar{S}_\Delta \sum_{m=0}^{A-1} \delta_m + (\varepsilon_1 - \varepsilon_2)$$

in which the sum of temporal discount factors operates as a scaling factor (see further below).

Equation (5) is a simplified version of a sibling difference model. While it is more restrictive than a complete VA model, specifications along this form are widely employed in the literature (e.g., Glewwe et al., 2001; Björklund and Sundström, 2006). A closely related specification is the sibling or family fixed effects model, which also is well-established within education research. For example, Deming (2009) employs family fixed effects models to investigate the impact of the Head Start program in the US on subsequent test scores. Similarly, Kerr et al. (2013) use an analogous specification to isolate the contribution of schooling reforms in Finland on later test score outcomes.

I implement the above model using the Uwezo data, placing children in each family in age order from oldest to youngest. The main school and individual inputs are selected to obtain

---

14 In the present case, estimates of a similar empirical expression using household fixed effects yields similar results to the difference model, which is expected given the dominant number of sibs included (of the correct age) in the sample in each multi-child household is two.
plausibly constant between-child differences, at least over recent periods. In a basic model, denoted specification A, I only include the highest school grade attended (representing $S$) and the child’s age (representing $C$). In an extended specification (denoted B), I add the child’s gender, whether she attends private school, whether she has presently dropped out, and proxies for individual ability, which are two dummy variables for whether the child correctly answers general knowledge ‘bonus’ questions. Specifications A and B use composite achievement measures as the outcomes of interest. However, as all unobserved aspects of individual ability will be relegated to the error term in these models, a further specification (denoted C) replaces composite achievement with the metric of numeracy achievement and adds a (composite) metric of literacy to the right-hand side of the model. This serves as an additional control for individual ability as well as other unobserved child-specific inputs. However, the current literacy score also is likely to control for contributions of (general) school inputs that affect both numeracy and literacy. Consequently, the corresponding school productivity estimates must be understood in a conditional sense and may well represent a lower bound. A rationale for this specification is that literacy is often considered a critical foundation for learning in other subjects but also is somewhat less sensitive to school inputs relative to numeracy (e.g., de Zeeuw et al., 2015).

As noted, an advantage of the sibling difference model (equation 5) is that it only requires data in cross-section. Given we have multiple survey rounds, the last challenge is how to combine information across countries and rounds. Regression estimates of the sibling difference model for a given country pooling all rounds would yield (weighted) mean estimates of $\gamma$ and $\alpha$. This would be informative but may give a false sense of coefficient homogeneity. An alternative is to allow the main coefficients of interest to vary by survey round. So, to verify parameter stability, which may be undermined by differences in test form functioning, I estimate separate coefficients in each round. In turn, this yields a set of estimates that can be used to indicate upper and lower bounds on the contribution of schooling to learning. This approach also is consistent with the assumptions of the non-pooled ridit and IRT metrics, where scores should be interpreted as relative to the distribution of achievement in each specific round.

4.3 Results

Summary regression results are set out in Table 5 for each of the five achievement metrics, showing selected coefficients. The main effects in the body of each panel of the table give average coefficients for the highest grade attended and the child’s age. These estimates only allow additive differences in scores in each round and do not adjust for parameter variation across rounds. As such, they correspond to pooled estimates of equation (5), adding round-specific fixed effects. To investigate parameter stability (uncertainty), the bottom rows of each panel report maximum and minimum estimates for the grade effect ($\gamma$), taken from the complete specification that allows such effects to vary across rounds. Specifically, I calculate the 95% confidence interval estimates for each
round-specific grade effect; and looking over all of these, I report the highest and lowest in the table.
Table 5: Summary regression results, by metric & country

<table>
<thead>
<tr>
<th></th>
<th>A (composite)</th>
<th>B (composite)</th>
<th>C (numeracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KE</td>
<td>TZ</td>
<td>UG</td>
</tr>
<tr>
<td>Highest grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.49***</td>
<td>7.51***</td>
<td>10.21***</td>
<td>9.44***</td>
</tr>
<tr>
<td>(0.12)</td>
<td>(0.13)</td>
<td>(0.14)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Age</td>
<td>3.69***</td>
<td>3.63***</td>
<td>2.17***</td>
</tr>
<tr>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>R²</td>
<td>0.55</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>Grade (min)</td>
<td>8.59</td>
<td>5.07</td>
<td>8.59</td>
</tr>
<tr>
<td>Grade (max)</td>
<td>12.87</td>
<td>9.03</td>
<td>12.10</td>
</tr>
<tr>
<td>Age</td>
<td>3.69</td>
<td>3.63</td>
<td>2.17</td>
</tr>
<tr>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.09)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>R²</td>
<td>0.55</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>Grade (min)</td>
<td>5.94</td>
<td>3.84</td>
<td>6.57</td>
</tr>
<tr>
<td>Grade (max)</td>
<td>9.17</td>
<td>7.13</td>
<td>9.32</td>
</tr>
<tr>
<td>Age</td>
<td>3.11</td>
<td>2.88</td>
<td>2.76</td>
</tr>
<tr>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>R²</td>
<td>0.55</td>
<td>0.39</td>
<td>0.45</td>
</tr>
<tr>
<td>Grade (min)</td>
<td>7.07</td>
<td>3.85</td>
<td>6.79</td>
</tr>
<tr>
<td>Grade (max)</td>
<td>10.52</td>
<td>7.17</td>
<td>9.50</td>
</tr>
<tr>
<td>Age</td>
<td>10.71</td>
<td>10.47</td>
<td>7.58</td>
</tr>
<tr>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>R²</td>
<td>0.55</td>
<td>0.39</td>
<td>0.45</td>
</tr>
<tr>
<td>Grade (min)</td>
<td>21.25</td>
<td>13.63</td>
<td>23.27</td>
</tr>
<tr>
<td>Grade (max)</td>
<td>32.63</td>
<td>25.33</td>
<td>32.79</td>
</tr>
<tr>
<td>Age</td>
<td>11.17</td>
<td>11.04</td>
<td>7.58</td>
</tr>
<tr>
<td>(0.27)</td>
<td>(0.31)</td>
<td>(0.26)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>R²</td>
<td>0.52</td>
<td>0.39</td>
<td>0.45</td>
</tr>
<tr>
<td>Grade (min)</td>
<td>24.49</td>
<td>13.47</td>
<td>24.01</td>
</tr>
<tr>
<td>Grade (max)</td>
<td>36.59</td>
<td>25.68</td>
<td>34.20</td>
</tr>
</tbody>
</table>

Note: Columns indicate specifications then countries; panels (a)–(e) indicate the achievement metric; all models are sibling differences; specifications A and B use the composite test score as the outcome, specification B adds additional background covariates to those shown; specification C uses the numeracy score as the outcome and adds the literacy score to the RHS; standard errors (in parentheses) are clustered at the village level; reported coefficients are pooled averages.

Source: own estimates.
Four main findings can be highlighted. First, the grade effects are positive and highly significant in all specifications and for all countries. Contrary to the most pessimistic views about schooling, which tend to suggest almost no learning takes place in low income country schooling systems, this demonstrates that schooling does contribute to material learning. However, while the metrics are not directly comparable across metrics, the effect magnitudes appear moderate. Based on the estimates that use the composite score and include a full set of controls (specification B), the percent correct score increases by approximately 9% among children in Kenya for each year of schooling (grade), versus 7% among children in Tanzania. Given the mean percent correct score among children starting school is below 30% in all countries, it follows that multiple years of schooling are needed to attain a high level of competency on the tests (ceteris paribus). The two ridit metrics point to a similar conclusion – children of starting school age also score around 30 ridit points and a year of schooling is associated with a gain of under 8 points. So, to reach a high percentile (>90) would take much more than two years of schooling. Thus, in reference to the external benchmark of curriculum expectations, these rates of learning appear insufficient.

The IRT estimates are reported in (approximately) effect size units and are highly similar regardless of the pooling assumption applied. For the non-pooled IRT score, a year of extra schooling boosts Kenyan children’s scores by 26 units or, roughly, a third of a standard deviation. This compares to 24 units in Uganda and 17 units in Tanzania, which is about a fifth of a standard deviation. In high income countries, a very crude rule of thumb is that children progress by about 50% of a standard deviation on (vertically-scaled) standardized tests for each additional year of schooling (e.g., OECD et al., 2015). While the Uwezo tests are not internationally comparable and may be biased due to floor/ceiling effects, the smaller effect sizes due to schooling in East Africa nonetheless corroborate a view that school productivity is frail.

Second, the results show that model specification matters. Moving across the three specifications from left to right in the table, the contribution of the grade effects tend to fall. This reflects the basic point that these effects become conditional on a more extensive set of controls. In the case of the final model, the outcome of interest is the numeracy score, which is specified as a function of the child’s literacy attainment. By construction, the grade effect estimates under this specification exclude any contribution of schooling to literacy that boosts maths attainment. The fact that we find positive and significant effects under this specification nevertheless supports the assertion that some learning is taking place in schools.

Third, if we follow conventional external benchmarking exercises and directly compare the magnitude of mean grade effects between countries, the country ranks are not entirely stable. This is shown in Table 6, which assigns a rank to each country based on the pooled grade effect measure for each specification and metric. For the ridit and IRT metrics, the shift from the pooled to the non-pooled scales is sufficient to switch ranks between Kenya and Uganda. So, the choice of metric and specification does seem to matter for country comparisons. However, coefficient estimates for Kenya and Uganda are frequently alike, and Tanzania shows consistently the lowest grade effect
magnitudes. In other words, similar qualitative patterns emerge across the metrics and models.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KE</td>
<td>TZ</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Note: ranks are derived from results shown in Table 5; KE is Kenya; TZ is Tanzania (mainland); and UG is Uganda. Source: own estimates.

Fourth, a focus on mean effects gives a false sense of precision. Additional sources of error, such as those associated with variation in test functioning over time, appear to be wide. This is indicated by the range between the maximum and minimum grade effects, which can be used to back-out an estimate of the (adjusted) standard error on $\gamma$. These results indicate that the adjusted errors are at least five times larger than the pooled regression-based standard errors reported in the table. So, there is significant uncertainty around the precise magnitude of grade effects. In turn, this substantially diminishes our ability to make precise comparative statements (e.g., bilateral comparisons). Indeed the overlap in confidence intervals between countries is so large that no single country dominates another in any specification.

As a final analytical contribution, it is appropriate to consider alternative benchmarks. Section 2.3 argued that the use of external benchmarks to evaluate the magnitude of grade effects may not always be very helpful, especially where achievement metrics are not strictly equivalent. An alternative basis for forming normative judgements comes from the the broader literature on interpreting effect sizes in empirical education research. In place of vague rules-of-thumb or absolute comparisons, Hill et al. (2008) argue that: “effect sizes should instead be interpreted with respect to empirical benchmarks that are relevant to the intervention, target population, and outcome measure being considered.” (p. 172). The same authors recommend various approaches to developing such benchmarks. One is to identify changes (effects) that might be expected to occur in the absence of any intervention. In the present case, it follows that grade effects in early primary school may be benchmarked against the background rates of cognitive growth that are observed in the same context but which are not driven by exposure to schooling. A natural candidate is the effect due to chronological age, conditional on other inputs. This also is attractive since this kind of internal benchmark can be established using estimates for $\alpha$ from equation (5). Thus, an alternative metric of school productivity is the relative grade effect, defined as: $\gamma^* = \gamma / (\gamma + \alpha)$, which gives the proportional contribution of a year of schooling to the total gain in achievement expected from completing one grade and becoming one year older. This estimate has the further advantage of

$15$ Given by: $\sigma_\gamma = (\gamma_{\text{max}} - \gamma_{\text{min}})/[2\Phi^{-1}(0.975)]$.

$16$ Hereafter, $\alpha$ refers only to the coefficient on age.
removing the unwanted scalar in equation (5). Since this scalar may vary across countries (and survey rounds) it provides a more rigorous basis for comparison across countries.

The strategy of evaluating grade effects against a plausible internal counterfactual is not new. As long as the internal benchmark is meaningful (and rigorous), estimates of this sort are intuitive and informative sui generis. Moreover, it may be more legitimate to compare relative effect sizes across different contexts as opposed to making absolute comparisons. Internally benchmarked effects, such as the relative grade effect, take into account differences in counterfactual learning gains and thereby (automatically) adjust for differences in test difficulty – i.e., for an ‘easy’ test the counterfactual learning gain would be larger. Furthermore, counterfactual rates of learning plausibly capture differences in contextual factors, such as childhood nutrition, which may have an interactive or conditioning effect on learning potential (in and out of school). As such, they go some way to ameliorate the difficulties of making bilateral comparisons where tests are not strictly comparable and contexts are distal. Furthermore, previous studies use the ratio or difference of estimated grade effects to estimated age effects to evaluate the performance of schools across different locations (Cahan and Cohen, 1989; Cliffordson, 2010; Lee and Fish, 2010). Although these studies show significant variation in the absolute size of both grade and age effects, their ratio is more stable and typically falls between one and two – equivalent to a relative grade effect of between 0.50 and 0.66. These magnitudes imply that schooling accounts for the majority of gains in achievement over a hypothetical year. However, in extremely challenging contexts, such as refugee camps, further evidence suggests that grade effects are substantially smaller in relation to age, implying little extra learning takes place in school (Jabr and Cahan, 2014).

The regression results of Table 5 include estimates of both age and grade effects. Thus, it is straightforward to calculate the relative grade effect as well as the corresponding maxima and minima ratios over assessment rounds (as before). These results are summarised in Table 7 and the distribution of survey round-indexed mean relative effects are shown in Figure A6(a). The relative grade effect ranges from a minimum of 33 percent to a maximum of 98 percent, with an overall mean of 72 percent. While these values span a wide range, the mean effects are remarkably consistent within countries, especially for each given outcome metric. More critically, the mean effects are remarkably consistent with – if not somewhat larger than – the grade-age effect ratios found in high income countries; see Figure A6(b). In other words, in keeping with evidence from high income countries, schooling is a dominant factor behind achievement gains in East Africa, contributing more than twice the gains due to ageing.

The important point here is that when we evaluate learning gains in light of background (counterfactual) rates of learning, the magnitude of achievement gains due to schooling observed in East Africa no longer appear quite so inadequate. The fact that assessments of school productivity differ so much between absolute and relative metrics (external versus internal benchmarks) in turn suggests

---

17 The relative grade effect is preferred to the direct grade/age ratio because the former is more stable when the age effect is very small.
that broader conditions affecting children’s capacity to learn may be (particularly) compromised in East Africa – i.e., low school productivity reflects more than just ‘bad’ schools. This connects with the literature on the critical importance of early childhood development for future achievement (e.g., Heckman, 2006; Cunha and Heckman, 2008). It also resonates with studies that emphasise schools cannot be viewed as technical units of human capital production that sit in isolation from the rest of society (Wagner, 2010; Alexander, 2012). Rather, learning must be understood and assessed relative to local constraints. From this perspective, the notion that schools in East Africa are performing far below what can be reasonably expected given local conditions becomes less tenable.

Table 7: Relative grade effects (regression-based)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Spec</th>
<th>KE</th>
<th>TZ</th>
<th>UG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Max</td>
</tr>
<tr>
<td>% correct</td>
<td>A</td>
<td>63.0</td>
<td>74.0</td>
<td>85.8</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>63.4</td>
<td>74.1</td>
<td>86.3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>37.8</td>
<td>74.1</td>
<td>74.1</td>
</tr>
<tr>
<td>Ridit (pooled)</td>
<td>A</td>
<td>60.0</td>
<td>70.6</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>58.9</td>
<td>70.4</td>
<td>82.8</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>33.0</td>
<td>70.4</td>
<td>71.1</td>
</tr>
<tr>
<td>Ridit (non-pooled)</td>
<td>A</td>
<td>63.1</td>
<td>73.3</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>63.4</td>
<td>73.4</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>47.9</td>
<td>73.4</td>
<td>76.3</td>
</tr>
<tr>
<td>IRT (pooled)</td>
<td>A</td>
<td>60.4</td>
<td>70.9</td>
<td>82.9</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>59.3</td>
<td>70.8</td>
<td>83.2</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>34.8</td>
<td>70.8</td>
<td>71.7</td>
</tr>
<tr>
<td>IRT (non-pooled)</td>
<td>A</td>
<td>62.4</td>
<td>72.6</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>62.6</td>
<td>72.6</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>46.5</td>
<td>72.6</td>
<td>75.2</td>
</tr>
</tbody>
</table>

Note: Cells show estimates of relative grade effects, calculated as the ratio of estimated grade effects to the effects due to ageing plus schooling; KE is Kenya; TZ is Tanzania (mainland); and UG is Uganda; pooled metrics are calculated using all rounds and countries simultaneously; non-pooled metrics use all rounds for each country separately; IRT and % correct scores are standardized to the distribution(s) of children who have never attended school.

Source: own estimates.
5 Conclusion

This paper began by acknowledging there are large gaps in academic achievement between children of the same age across the world – i.e., we face something of a ‘global learning crisis’. A widespread and seemingly natural policy diagnostic associated with this observation is that school productivity is inadequate in many developing countries. More simply, children are not learning enough in school. The overall aim of this paper was to examine whether this diagnosis holds up to scrutiny. That is, do schools in low income countries really produce so little learning?

My answer to this question was split in two parts. The first part evaluated existing evidence on school productivity in low income contexts. To start, I reviewed the inherent technical challenges involved in undertaking normative assessments of this kind. Three main challenges were identified – establishing a suitable metric of achievement (i.e., one that is valid and reliable); estimating a suitable model of learning, so as to isolate the unique contribution of schooling to achievement gains; and applying a meaningful meter or benchmark, so as to distinguish between good and bad performance. It was shown that these challenges should not be under-estimated. Even the most sophisticated and high-profile national and international learning assessments face ongoing technical critiques. In this vein, it is germane to recall the note of warning signalled by Ballou (2009): “many social scientists using test scores to evaluate educational institutions and policies have little or no training in measurement theory” (p. 352). Since the technical issues involved in normatively evaluating school productivity are both extensive and controversial, this review underlines that a great deal of caution is required before strong, policy-relevant conclusions are drawn.

The above point echoes existing concerns around school accountability measures used in high-resource settings (e.g., Kane and Staiger, 2002; Guarino et al., 2015; American Educational Research Association, 2015). When applied to low income settings they attain greater force. Indeed, learning assessments undertaken in low income countries have generally been designed to monitor overall achievement levels (over time). However, they have not been designed to isolate the unique contribution of specific inputs, such as schools. As such, many existing claims about school productivity in low income countries fail to isolate the unique contribution of schools to differences in achievement. Furthermore, even while achievement gaps may provide a logical basis to argue that schools should do more, this argument is mute on whether actual rates of school productivity are low(er) in low-achieving countries. It is also mute on whether it is reasonable, given existing socio-economic conditions, to expect that schools alone can significantly close the achievement gaps. This raises questions about what constitutes a reasonable or meaningful benchmark to evaluate performance.

The second part of the paper (Section 4) moved to the specific case of East Africa. Using data collected by the Uwezo initative since 2010, I investigated what can be said about school productiv-
ity in these countries. To begin, I showed that data quality concerns limit the nature and possible robustness of achievement estimates. In particular, there are prevalent floor and ceiling effects and it is not clear that the tests are directly comparable, even within countries over time. Cognizant of these drawbacks, I set out a sibling difference model that can be feasibly estimated using the Uwezo data. Moreover, to address concerns about test (in)equivalence, I proposed that broad bounds on the estimated grade (and age) effects should be derived from coefficient estimates indexed to each survey round.

Turning to the results for East Africa, and in keeping with previous literature (Seltzer et al., 1994; Briggs and Weeks, 2009; Bond and Lang, 2013), I confirmed that technical choices do matter – e.g., for measuring achievement gaps, as well as for ranking countries in terms of school productivity. Furthermore, when uncertainty about test comparability is accounted for, we cannot easily distinguish between countries in terms of the precise magnitude of estimated grade effects (school productivity). Even so, a consistent qualitative pattern of findings emerged. On the one hand, when applying the external benchmark of national curriculum expectations, I found it would take much more than two years of schooling for the average child to achieve the full set of competencies expected at the grade 2 level. Thus, from an absolute perspective, school productivity in East Africa appears inadequate.

On the other hand, internal benchmarks developed from counterfactual or background rates of cognitive development – namely, the conditional age effect – indicated that the relative contribution of schooling is not so different from that in developed countries. Various studies, using different learning assessments, show that the ratio of grade effects to age effects is approximately two (or lower) in high income countries. In East Africa, multiple achievement metrics from the Uwezo data indicate the interquartile range of the same ratio runs from 1.9 to 2.8 (see Figure A6b). As such, this study has highlighted a large difference between absolute and relative metrics of school productivity.

One interpretation of these findings is that children’s learning potential in East Africa is materially conditioned by a range of background socio-economic factors, such as early childhood nutrition. This account connects to a large body of other evidence that indicates public policies to promote early childhood development and school-readiness arguably may have a far higher long-run payoff in terms of education and labour market outcomes than measures aimed at schools alone. So, to rephrase Pritchett’s (2013) mantra, we are wise to remember that ‘learning ain’t just about schooling’. The proposed metric of relative school productivity provides one starting point to evaluate school performance in light of the broader set of conditions that structure learning potential. From this angle, schools in East Africa may not be doing so poor a job of raising achievement. Of course, this performance is hardly satisfactory from the absolute perspective of closing international achievement gaps; but it is understandable. This argument may be provocative. However, instead of demonizing schools and teachers for their failures, it provides an entry point for more constructive dialogue and encourages education policy to be re-embedded in its local context.
References


Figure A1: Observed vs. true learning profiles in presence of floor effects

<table>
<thead>
<tr>
<th>Score</th>
<th>Grade</th>
<th>Country A</th>
<th>Country B</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

Note: observed learning profile is subject to floor effects (at zero), which are substantially larger in country B; also see text.

Source: own calculations.
Figure A2: DIF analysis of English reading and numeracy tests, across countries (2013)

(a) English

(b) Numeracy

Note: .
Source: own calculations.
Figure A3: DIF analysis of numeracy and literacy test items between rounds, Kenya

(a) English

(b) Numeracy

Note: .
Source: own calculations.
Figure A4: DIF analysis of numeracy and literacy test items between rounds, Tanzania

(a) English

(b) Numeracy

Note: .
Source: own calculations.
Figure A5: DIF analysis of numeracy and literacy test items between rounds, Uganda

(a) English

(b) Numeracy

Note: .
Source: own calculations.
Figure A6: Internally benchmarked school productivity measures

(a) Relative grade effect

(b) Grade / Age effect

Note: figure gives box and whisper plots for (a) the relative grade effect: $\gamma_t / (\gamma_t + \alpha_t)$; and (b) the ratio of estimated grade to age effects: $\gamma_t / \alpha_t$, indexed by survey round, across countries.

Source: own calculations.