Learning More With Every Year:
School Year Productivity and International Learning Divergence*

Abhijeet Singh
University of Oxford†

April 1, 2015

Abstract

How important is the differential productivity of primary schooling across countries in explaining international differences in human capital? In this paper, I present the first micro-econometric evidence on this issue using a unique child-level panel dataset with linked assessments of quantitative ability at 5 and 8 years of age from four developing countries – Ethiopia, India, Peru and Vietnam. I first document that although some cross-country gaps in test scores are evident already at 5 years of age, prior to school entry, these gaps grow substantially in the first years of schooling: children in Vietnam consistently score highest on average followed by Peru, India and Ethiopia respectively. Using value-added estimates, together with an RD/IV approach using discontinuities in grade completion arising due to month of birth and country-specific enrolment guidelines, I document sizable cross-country differences in the productivity of a school year. School year productivity, measured as learning-gains-per-grade-completed, varies from 0.45 SD-per-year in Vietnam to, for example, 0.2 SD-per-year in Peru. This differential productivity of a year in primary school, rather than differences in child endowments, nutrition, background characteristics or time use, largely explains the early divergence between Vietnam and other countries: equalizing this parameter to Vietnamese levels, leaving all endowments including learning at 5 years unchanged, closes the entire achievement gap between Vietnam and Peru, and 70% of the gap between Vietnam and India, at 8 years of age.

*I am grateful to my doctoral supervisors, Stefan Dercon and Albert Park, and to Steve Bond and John Muellbauer for encouragement and feedback on this paper and other parts of my doctoral research. I am also very grateful to Harold Alderman, Orazio Attanasio, Jere Behrman, Santiago Cueto, Le Thuc Duc, James Fenske, Doug Gollin, Paul Glewwe, Michael Keane, Karthik Muralidharan, David Lagakos, Lant Pritchett, Caine Rolleston, Justin Sandefur, Todd Schoellman, Simon Quinn, Nicolas Van de Sijpe, Adrian Wood and various seminar participants at Oxford, NEUDC 2014 (Boston University), Sheffield, Goettingen and Indian Statistical Institute, Delhi for providing invaluable comments on previous drafts. I thank the Young Lives Project for providing the data in this paper and supporting this study. Young Lives is core-funded by UK aid from the Department for International Development and was co-funded from 2010 to 2014 by the Netherlands Ministry of Foreign Affairs.

†Address: Nuffield College, University of Oxford, 1 New Road, Oxford OX1 1NF, UK. Email: abhijeet.singh@qeh.ox.ac.uk

1 Introduction

The role of human capital in explaining differences in economic growth across countries has been of central interest to economists. Early studies, such as Mankiw, Romer and Weil (1992), emphasized the differential quantity of human capital as measured by years of education and the subsequent impacts on growth. More recent work recognizes additionally that years of schooling mask differences in the *quality* of human capital, which varies importantly across countries, and can be quantitatively more important in explaining variations in economic growth.[1]

Two key questions, however, still remain unanswered: at what ages do these gaps in human capital emerge across countries and what is the role of the productivity of schooling, rather than merely the quantity of schooling, in the production of these international differences in human capital? These questions relate centrally both to the macroeconomic literature on growth accounting mentioned above and to a large microeconomic literature seeking to identify the age patterns in the evolution of learning gaps and the domains and ages in which interventions may be most required.[2]

In this paper, I address these questions using a unique child-level panel dataset which tracks two cohorts of children between 2002 and 2009 from four developing countries – Ethiopia, India, Peru and Vietnam. Specifically, I provide the first internationally comparative analysis of educational trajectories on linked tests at preschool and primary school ages of 5 and 8 years, a critical period of childhood for skill formation; present comparable causal estimates of the productivity of school systems (measured as the learning-gains-per-grade-completed) in these four countries; and examine the contribution of the differential productivity in explaining international differences in the level of human capital (measured directly by comparable test scores).

The data are particularly suitable for this analysis. Most importantly, the surveys administered the same tests at each age across countries: thus, I can generate comparable distributions of test scores for quantitative proficiency across countries at both 5 and 8 years of age. Further, the child-level panel structure of the data, combined with detailed background information, allows me to analyze learning dynamics and estimate rich production functions of achievement. Since the data were collected through home

[1] This differential quality has been measured both directly through test scores on international learning programs (see Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008, 2012a,b; Kaarsen, 2014) and indirectly through the returns to education of immigrants from different countries in the US labour market (see, e.g., Hendricks, 2002 and Schoellman, 2012). For results on the contribution of differential quality of schooling to growth see e.g. Barro (2001), Erosa et al. (2010) and Manuelli and Seshadri (2014).

[2] For example, Fryer and Levitt (2004, 2006) document that it is precisely in the first years of schooling that the Black-White test score gap in the US substantially expands, indicating that this is perhaps a critical period for intervention. Such understanding is especially important in light of a now-large literature which documents that the effectiveness of interventions varies over the age of children (see Heckman and Mosso, 2014 for a recent review). In a recent macroeconomic contribution, Manuelli and Seshadri (2014) also pay explicit attention to differences in human capital arising in early childhood as distinct from differences arising in schooling or through post-school training.

visits of a random sample of children in given birth cohorts, they do not suffer from selection issues arising from non-enrolment or non-attendance that are characteristic of school-based assessments in developing countries. The data are also particularly well suited to this analysis since the four countries display remarkably different levels of achievement.[3]

Using these data, I first document that although gaps across countries are already evident by 5 years of age, before children have entered school, they expand substantially by 8 years of age in the first years of primary schooling. Further, I use the individual-level panel dimension of the data, combined with household-level information, to estimate value-added models of achievement production. Specifically, I focus on assessing whether test score divergence across countries is explained by differing child-level endowments and home investments, differing time use patterns, and differing exposure to and effectiveness of a grade of schooling in the four countries.

The central result of this paper is that the *productivity of a school year*, measured as the 'learning-gains-per-grade-completed' in value-added models, varies widely across countries and accounts for most of the gap observed between the learning levels of children in Vietnam, the only 'high-performer' in our sample, and the other countries. Estimates of the productivity of a school year are statistically indistinguishable between Ethiopia, India and Peru but significantly higher in Vietnam. Equalizing this parameter to Vietnamese levels closes the entire gap at 8 years of age between Vietnam and Peru, and 70% of the gap with India, even keeping all other inputs including learning levels at 5 years at their initial levels. For two countries, Peru and Vietnam, I validate the value-added estimates of productivity using a regression discontinuity (RD)/instrumental variable (IV) approach. I use discontinuities in the number of grades completed arising from enrolment thresholds and child month-of-birth for the identification of grade productivity and do not find any significant evidence of systematic bias in the value-added estimates compared to instrumental variable estimates based on this discontinuity. Preferred estimates suggest that whereas an additional grade completed in Vietnam leads to a causal increment of about 0.45 SD in mathematics test scores, the effect is only about 0.2 SD in Peru.

This paper makes three key contributions. First, it presents the first analysis of the evolution of gaps in cognitive achievement across countries in the early years of education, a critical period for the divergence in test scores, using internationally comparable child-level panel data. While similar exercises have previously been carried out within the context of individual developed countries while studying socioeconomic or racial gaps

[3]Math scores of 15-year-old students in the 2012 PISA assessment differed by 1.4 standard deviations (SD) between Vietnam and Peru, two of the countries covered in this paper – while Vietnam scored higher than the US and the UK, Peru was the last of 65 countries covered in the assessment. In comparison, the difference between the US and Finland, the highest scoring Scandinavian country, was 0.38 SD(OECD, 2013). India did not participate in the 2012 PISA round but had been surveyed by PISA in a follow-up to the 2009 round: students in these states scored lower than Peru by about 0.15-0.3 SD (Walker, 2011). Ethiopia has not been covered by international assessments thus far; in our data, the average performance in the Ethiopian sample consistently lies below that in the other countries.

in test scores (see e.g. the voluminous literature on the Black-White test score gap in the US), I am not aware of any research that attempts to link analyses comparably across countries. Second, it provides the first comparable causal estimates of school productivity across countries, which are identified using individual-level panel data combined with quasi-experimental variation, and quantifies the role of differential schooling productivity in explaining learning gaps using microeconometric production function estimates. Third, in validating results from value-added models using quasi-experimental variation, the paper also adds to a stream of recent methodological work which seeks to compare value-added estimates to estimates using experimental or quasi-experimental variation (see e.g. Chetty et al., 2014; Andrabi et al., 2011; Deming et al., 2014; Singh, 2015).

The analysis and results presented here relate to several strands of the literature within economics. Methodologically, this paper relates closely to the literature on the emergence and evolution of test score gaps between different racial groups or gender (see e.g. Fryer and Levitt, 2004, 2006, 2010, 2013; Todd and Wolpin, 2007), on value-added models of achievement using household-based panel data (see e.g. Todd and Wolpin, 2003, 2007; Fiorini and Keane, 2014) and on mapping test scores on a comparable metric using Item Response Theory (IRT) models.[4]

Substantively, in focusing primarily on the gaps across countries in the stock of human capital (learning levels), and in the productivity of a grade of schooling, this paper has direct relevance for two strands of the applied macroeconomics literature — studies seeking to explain the effect of differential human capital and schooling quality to growth (Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008; Schoellman, 2012; Kaarsen, 2014); and studies seeking to study differences in productivity across countries in different sectors using micro data.[5] Results on differential grade productivity may also indirectly relate to the measured differences in the quality of management across different countries (Bloom and Van Reenen, 2007, 2010), including in schools (Bloom et. al., forthcoming).

Finally, in focusing on patterns and determinants of skill acquisition at the child level, the analysis also relates directly to a large and rapidly growing microeconometric literature on the economics of education in developing countries that documents low levels of learning (see Glewwe and Kremer, 2006 for an authoritative review), studies trajectories in skill acquisition (see e.g. Schady et. al., forthcoming), and experiments with different interventions to improve these low levels of learning (see Kremer et al. 2013 and McEwan 2013 for meta-analyses).

[4]These models are commonly used in comparative educational assessments (see e.g. Van der Linden and Hambleton, 1997; Mullis et al., 2004; OECD, 2013) but are rare within development economics (for notable exceptions, see Das and Zajonc (2010) and Andrabi et al. (2011)).

[5]See, for example, Hsieh and Klenow (2009) who study factor productivity and misallocation across Chinese and Indian firms or Gollin et al. (2014) who investigate the agricultural productivity gap in developing countries using household survey data.

The rest of this paper is structured as follows: Section 2 describes the data used in this paper; Section 3 investigates stochastic dominance of test outcomes across age groups and assesses whether learning gaps seems to narrow or widen with age between countries; Section 4 estimates value-added models of achievement, presenting assessments of school-year productivity, and validates these estimates using the RD/IV approach; it also presents further sensitivity analyses for robustness of results; Section 5 discusses the findings and concludes.

2 Data and context

2.1 Data

This paper uses data collected by the Young Lives study in Ethiopia, India (Andhra Pradesh state), Peru and Vietnam which has tracked two cohorts of children over multiple rounds since 2002. The older cohort (born in 1994/95) was aged about 8 years, and the younger cohort (born in 2001/2) between 6-18 months, at the time of the first wave of the survey in 2002. In each country, about 2000 children of the younger cohort and 1000 children in the older cohort were surveyed. Two subsequent waves of household-based data collection were carried out in 2006 and 2009. The data are clustered and cover 20 sites in each country across rural and urban areas.[6] In cases where children have moved from the original communities they were surveyed in since 2002, the study tracked them to their new location. As a result attrition in the data is very low with about 95% of the original sample in the younger cohort, the main focus of analysis in this paper, still in the survey in 2009 in each country. Figure 1 presents the age of children in the two cohorts at each survey wave[7].

In each round, a range of background information and child-specific data, including nutritional outcomes and academic achievement, was collected.[8] Quantitative skills were tested for the younger cohort in 2006 (then aged about 5 years) using the 15-item Cognitive Developmental Assessment tool developed by the International Evaluation Association Preprimary Project and in 2009 through a 29-item mathematics test. The survey instruments, including tests, were harmonized across countries in each round.[9]

[6]Sites correspond to sub-districts in Ethiopia (kebeles), India (mandals) and Vietnam (communes) and to districts in Peru. Sites were chosen purposively to reflect the diverse socio-economic conditions within the study countries and therefore are not statistically representative for the country: comparisons with representative datasets like the DHS samples do show however that in each of the countries, the data contain a similar range of variation as nationally representative datasets (Outes-Leon and Sanchez, 2008; Kumra, 2008; Escobal and Flores, 2008; Nguyen, 2008).

[7]Fieldwork typically took between 4-6 months in each country in each round. The timing of the survey rounds shown in Figure is thus only indicative.

[8]Summary statistics on these variables will be presented in Section 4 at the point they are being introduced in value-added specifications of achievement production.

[9]See Cueto et al. (2009) and Cueto and Leon (2013) for details of the psychometric testing in Young Lives across different rounds.

In this paper, I focus entirely on results from the younger cohort of children who were aged about 5 years in 2006 and 8 years in 2009. This is both because I wish to particularly focus attention on the early divergence in test scores in primary school, but also because of an important limitation in the older cohort: since enrolment thresholds and month-of-birth are not sufficiently predictive of grade completion by 15 years of age, I cannot similarly generate RD/IV estimates of grade productivity. Moreover, there are additional concerns caused by details of the testing in the older cohort – in particular, because the tests in Young Lives are focused on relatively simple competences, they are much less aligned to the skills being taught to children in secondary schools.[10]

Table 1 presents descriptive information about the educational trajectories and progression of children in the Young Lives sample at different ages which will be of central importance in interpreting all results in this paper. Two patterns from this Table are worth highlighting: ever-enrolment is high across all countries with nearly all children having been enrolled at some point in primary school; and, second, the age of entry varies importantly between countries, being the lowest in India (with about 44% of children already in school by 5 years of age) and highest in Ethiopia (with a significant proportion of the sample yet-to-enrol at 8).[11]

The analysis in this paper requires generating comparable test scores across countries at 5 and 8 years. This is done using Item Response Theory (IRT) models which provide several advantages for the purpose of this analysis: first, by explicitly mapping the relationship between the probability of answering a particular test item correctly and an individual's ability, they provide a less arbitrary aggregate measure of proficiency than a percentage correct score which assigns equal weight to all questions, regardless of difficulty; second, they allow for comparable linking across samples; and third, they provide for a more robust framework for diagnosing comparability in item performance across contexts, which may be violated due to, for example, translation issues or cultural specificity of items.[12] The models are estimated as in Das and Zajonc (2010) who linked responses to mathematics questions administered in two states of India to the TIMSS 8th Grade test.

Item Response models only identify ability (θ) up to a linear transformation i.e. any transformation of the form $a + b\theta$ is an equally valid test score. This implies that, in the absence of common items which can be used to 'anchor' estimates in two samples,

[10]Interested readers may, however, wish to consult the Working Paper version of this paper (Singh, 2014) which additionally presents both descriptive results on learning divergence and value-added models of learning achievement for the older cohort between 12–15 years and discusses these data issues in greater detail. In addition, analysis for the older cohort presented there links test scores to the TIMSS 4th grade assessment (2003 round) which was administered in 29 countries.

[11]About 23% of the sample is out of school in Ethiopia even at the age of 8 years, which reflect the much later age for starting school in Ethiopia (and the much greater dispersion in this starting age). Most non-enrolled children observed at the age of 8 years of age in 2009 are expected to join schools in later years; in the older cohort, for example, 95% of the children in the Ethiopian sample was enrolled at the age of 12 in 2006 in the same communities.

[12]The use of IRT models is standard in educational assessments to generate test scores that are comparable over time or across different populations; it is used, among other applications, in the generation of test scores for the GRE, SAT, TIMSS, PISA and NAEP in the US.

we cannot compare the absolute levels of achievement across two samples. Since tests were not harmonized across rounds, I cannot link test scores on a comparable scale over time. Appendix A provides a brief explanation of IRT and details the procedures for the generation of test scores, as well as analyses to check for and accommodate instances of the differential performance of items across the four countries.

3 When do gaps emerge and how do they evolve?

In this section, I analyse whether a clear ranking of country samples is discernible at these early ages (5 and 8), whether there are systematic differences in the level of achievement between children in different countries, and whether there are additionally differences also in how much they learn over time.

Table 2 presents descriptive statistics of the quantitative ability scores in the two rounds of the survey. The test scores are comparable across countries at each age and are normalized internally to have a mean of 500 and a standard deviation of 100 at each age in the pooled sample; test scores are not linked over time and therefore cannot be directly compared across age groups.

A reasonably clear ranking is already evident at the age of 5 years, which pre-dates schooling for most of the sample: Vietnam and Peru at the top-end, followed by India and Ethiopia respectively.[13] It is notable though, given large differences at later ages including the PISA assessment at 15 years, that quantitative achievement in Vietnam and Peru does not differ significantly at preschool ages. The second important pattern is that the ranking of countries in the Young Lives sample is mostly unchanged across age groups - although a clear gap has opened up between Vietnam and Peru, with India and Ethiopia following, the general ranking of the four country samples is remarkably stable. This ranking is also the same as the ranking implied by the PISA test scores at 15 for Vietnam, Peru and India (see footnote 3).

Conclusions formed on the basis of the mean comparisons also hold true across the entire distribution with a clear pattern of first-order stochastic dominance. The CDFs of the estimated test scores are plotted for both the age groups in Figure 2. This is important because mean comparisons are not adequate to make judgments about the entire distribution of learning across countries. As Bond and Lang (2013) point out, citing Spencer (1983), the ordinality and arbitrary normalization of test scores implies that the only way of reliably ranking samples is to look at the cumulative distribution functions (CDFs) of achievement.

[13] Given the literature from OECD countries about early emergence of cognitive gaps, this is perhaps unsurprising. However, this pattern is notable because analysis on the cognitive impact of environmental influences which precede school enrolment (with the exception of nutrition or health shocks, (e.g. Glewwe et al., 2001 and Maccini and Yang, 2009) remains very limited in developing countries. For notable exceptions using Latin American data see Berlinski et al. (2008) and Berlinski et al. (2009). There is also a broader inter-disciplinary literature (see e.g. Engle et al. 2007) but it is mostly associational.

Since test scores are not linked across age groups, Figure 2 cannot show whether inter-country gaps become larger with age. Further, even if gaps have grown, the Figure cannot answer whether any further divergence is only caused by amplification of initial gaps (through the self-productivity of skills) or through other channels in achievement production. In Figure 3, I present non-parametric plots of achievement in 2006 and achievement in 2009 for the four country samples. The essential idea behind these graphs is simple: conditional on test scores in 2006, do we see children in the four countries achieve similar results in 2009 (in which case gaps at later ages only reflect past divergence), or do we see children in some countries perform better than children in other countries who had scored similarly in 2006 (in which case there is additional divergence)?

There seems to be considerable divergence between countries between 5 and 8 which is similar to the ranking of the country samples on the levels of achievement at age 5: children in Vietnam learn more than children in Peru, who in turn learn more than children in India and Ethiopia respectively, even conditional on having achieved the same score at age 5.[14] The difference between the countries seems to be a difference in the intercepts and not the slopes which suggests that divergence between ages is not due to the initial learning levels in the four samples.

4 Sources of divergence: Results from value-added models

Analysis presented in the previous section is only partly informative about the sources of this divergence. Documenting patterns of learning even conditional on past achievement is insufficient by itself to say, for example, whether the divergence is primarily a factor of school inputs or a result of constant application of superior home inputs at every life stage in some contexts than others; from a policy perspective, however, identifying sources of divergence is of considerable interest. In this section, I estimate value-added models of achievement production to address this issue.

4.1 Do child-specific endowments explain divergence?

As a benchmark case, I first explore sources of achievement across the four countries at each age group as follows:

$$Y_{ic,a} = \alpha + \beta_1.\theta_c \quad (1)$$

[14] See also Rolleston (2014) and Rolleston et al. (2013) who document similar descriptive patterns of divergence in these data. The most important difference between their analysis and this paper is in the use of IRT models here which offer a better conceptual basis for cross-cultural comparison, allow different test items to contribute differently to the aggregate test score and provide a more continuous measure of ability in comparison to percentage correct scores. Rolleston (2014) and Rolleston et al. (2013) do not attempt an analysis of the sources of divergence across countries or the productivity of schooling.

$$+\beta_2.Y_{ic,a-1} \quad (2)$$

$$+\beta_3.X_{ic} \quad (3)$$

$$+\beta_4.TU_{ic,a} + \epsilon_{ica} \quad (4)$$

where $Y_{ic,a}$ is the test score of child i in country c at age a; θ_c is a vector of country dummy variables (with Ethiopia as the omitted category); X_{ic} is a vector of child-specific characteristics which includes the primary caregiver's education (in completed years), child's age in months at time of testing, child's height-for-age z-scores at time of testing (based on WHO 2005 standards), a wealth index based on durables owned by household and access to services, and dummy variables for being male and being the eldest child; TU is a vector controlling for time use across different tasks on a typical day (with sleeping being the omitted activity). The estimation is carried out by pooling data from all country samples.

Inclusion of controls is sequential as detailed in Equations 1-4 and naturally changes the interpretation of coefficients. Specification (1) displays mean difference between countries at the ages of 8 and 15 years; Specification (2) is the linear regression analogue of Figure 3 and shows the divergence between the countries, conditional on lagged individual test scores; Specifications (3) and (4) further explore if the divergence is explained by the levels of covariates in X or in current time use respectively. Time use is entered in the final step since it potentially conflates (through the categories of time spent at school or studying after school) school inputs with home-based inputs.[15]

Estimating achievement production functions is difficult as the full history of inputs applied at each age, as well as the full vector of child-specific endowments, is not observed in any dataset. Specifications (3) and (4), which include previous test scores and a range of controls, are commonly known as 'lagged value-added models' (VAMs) of achievement production(Andrabi et. al., 2011; Todd and Wolpin, 2007). These models attempt to deal with this problem by entering the lagged achievement score in the estimation as a summary statistic for child-specific endowments and the full history of inputs. These VAMs will be used as the workhorse specifications for the analysis of divergence in this paper although I will investigate possibilities of bias due to unobserved heterogeneity and measurement error later in this section which, as Todd and Wolpin (2003, 2007) discuss, may importantly bias value-added (VA) estimates.[16]

[15]The inclusion of time use categories should thus be considered here in the spirit of a bounding exercise, exploring the upper bounds of how much may be explored by means of variables determined at home. Given that categories of time use (e.g. time spent studying after school) are likely to be correlated with unobserved time-varying investments into children's learning (e.g. parental attention to schooling), coefficients on time use categories should be interpreted with care.

[16]It is relevant to note here that a large recent literature finds the observed level of bias in VA estimates to be low across a range of applications including in comparisons with experimental estimates (Kane and Staiger, 2008; Kane et al., 2013; Deming et al., 2014; Angrist et al., 2013; Singh, 2015), with quasi-experimental estimates (Chetty et al., 2014), dynamic panel data estimates (Andrabi et al., 2011) and in simulated data with a variety of non-random assignment mechanisms(Guarino et al., 2015). My

Summary statistics of the controls used in the estimation of achievement production functions are presented in Table 3. Results from the estimation of Specifications 1-4 are presented in Table 4. As is evident, the mean test scores are statistically significantly different across the four countries and substantial in magnitude (Col. 1). Controlling for the lagged test achievement (Col. 2) reduces the gap between countries somewhat. Gaps decline further upon inclusion of background variables in the input vector X but the magnitude of decline is small as a proportion of the initial gap. Inclusion of time use inputs has important effects, especially for Ethiopia, where the gap between Ethiopia and the other countries reduces substantially (especially with India where it is now at about a fifth of the initial cross-sectional gap). This pattern is likely a product of the enrolment profiles across country samples since a large proportion of 8-year old children in Ethiopia have not yet joined school.

The central pattern in Table 4 is that a substantial gap remains between the country samples; even in the most extensive specification, Ethiopia and Vietnam differ by between 0.9 SD at 8 years of age, accounting for more than 60% of the cross-sectional gap in test scores. It is only between Ethiopia and India that input levels seem to explain a substantial portion of the learning divergence between countries. These results suggest that while differences in endowments and socio-economic background play a role in creating differences across samples, it appears unlikely that this is the sole, or perhaps even the main, cause of divergence.

4.2 Does differential productivity of home inputs explain divergence?

Specifications 2-4 impose a strong assumption of common parameter coefficients on inputs across countries. This assumption is unlikely to hold; there is no reason to assume, for example, that a year of maternal education has an identical impact on child test scores in Vietnam and Ethiopia. In order to allow for maximum heterogeneity across the four countries, I estimate specifications (3) and (4) separately for each country thus allowing all input coefficients to differ. Results are presented in Table 5.[17]

As can be seen, the coefficients on specific inputs differ across countries. It is, however, difficult to directly read from this table the importance of this differing productivity for explaining test score gaps. In order to facilitate such comparison, I present some counterfactual examples which predict the mean level of achievement in each country sample by keeping the country's level of inputs fixed but varying the coefficients of the inputs (including the constant term) to match other countries.[18] The results are shown

results in this paper, comparing VA estimates to those based on a fuzzy regression discontinuity design and unable to reject equality, further add to this literature.

[17] In this table, and all other regressions estimates spearately at country-level, I cluster the standard errors by the survey site that the child was first observed in.

[18] These are two polar cases where I change either all inputs or all coefficients. In practice, from a policy perspective, it may not be feasible or even desirable to change all inputs or coefficients. Nor is

in Table 6 from specifications both with and without time use: the primary diagonal in each case, in bold, shows the actual level of achievement in the relevant country sample.

The results are informative. The difference between the average levels of achievement between Ethiopia and India, as also evident from Table 4, seems mostly a difference in the inputs of the children (including their test scores at age 5 which reflect investments in early childhood): equalizing these in the specifications with time use, but maintaining the same production function parameters as estimated in Table 5 for the country, reduces almost the entire gap between Ethiopia and India.[19] However, the difference between Vietnam, the only 'high performer' in our sample, and the other three countries seems to lie not in the endowments, including what children had learnt prior to school entry, but in the higher rates of learning afterwards: for each of the other three countries, considerably more of the learning gap is closed by equalizing productivity of inputs than by equalizing input levels. For example, both in specifications with and without time use, while raising Peruvian inputs to Vietnamese levels does not change mean achievement levels in Peru at all, equalizing the productivity of the inputs closes almost the entire gap; similarly, in the specifications with time use, the whole gap between India and Vietnam is covered by equalizing the productivity of inputs to Vietnam.[20]

It is clear from this exercise is that, whereas child endowments and the investments made in early childhood are undeniably important, the major divergence with Vietnam is after the age of 5 years. Considerably less clear is the source of this divergence. The major difference in productivity between Vietnam and the rest comes from the difference in the coefficients on the age in months in Table 5: we know that Vietnamese children seem to be learning more as they age each month than the children in the other countries but we don't quite know why.[21] This will be investigated in the later subsections.

choice limited to only choosing to shift elements of only one or the other vector; many more combinations could be explored. The purpose here is only to highlight two contrasting possibilities to assess the relative importance of these two channels (differences in the level of inputs and differences in input productivity) in explaining divergence.

[19]To see this, note that fixing inputs to Indian levels, in the specifictions with time use, produces a predicted mean level of 485 using Ethiopian coefficients and 495 using Indian coefficients. The difference between the two numbers is not statistically significant and we cannot reject that the full gap is thus closed. The reliance on the specifications incorporating time use is particularly relevant here since it captures the differences in enrolment between Ethiopia and the other countries at 8 years of age.

[20]To see this note that, in the specifications with time use variables, applying Vietnamese coefficients to Peruvian inputs shifts the Peruvian mean score from 516 to 557 but shifting inputs (while keeping coefficients unchanged) does not move scores at all. Similarly, applying Vietnamese coefficients to Indian inputs raises Indian mean scores from 496 to 563 which is near-identical to the Vietnamese mean; the reverse case (keeping Indian coefficients but applying Vietnamese inputs) reduces achievement, albeit not significantly.

[21]The difference between the coefficient on age in Vietnam and in the other countries is invariably statistically significant in cross-equation tests at conventional levels of significance.

4.3 Do differential exposure to schooling and differential productivity of schooling explain divergence?

As Table 6 documents, only a small portion of the divergence across countries till the age of 8 years, especially with Vietnam, is accounted for by the levels and differential productivity of observed home inputs across the four country samples; most difference seems to be an unexplained difference in trends. One possibility that may account for divergence after 5 is the differential exposure to schooling across the four country samples; for example, Ethiopian children enter school much later than in the other countries (Table 1) and thus have less schooling at every age in the sample than the other countries. Similarly we would expect given previous work that the quality of schooling differs across these contexts, which could also contribute to the growth of these gaps.

In order to study the importance of these schooling-based sources of divergence, I estimate the following specifications:

$$\begin{aligned} Y_{ic,a} &= \alpha_c + \beta_1.Y_{ic,a-1} + \beta_2.X_{ic} + \beta_3.grade_{ica} && (5) \\ &\quad +\beta_4.TU_{ic,a} + \epsilon_{ica} && (6) \end{aligned}$$

where in addition to variables defined previously, I also include a variable for the highest grade completed by the child at age *a*. As in the previous specification, the estimation is carried out separately for each country sample and I estimate the production function both with and without the time use inputs. The parameter of interest is β_3 which, if it differs across countries, would indicate differences in the amount of progress in quantitative skills per grade completed across the different educational systems.

Note that grades completed may be regarded as an *outcome* of educational systems rather than merely an input into learning and thus raise concerns about its endogeneity (for example if, as is likely, the same factors determine both grade completion and the amount learnt in school). Identification of β_3 in this case rests on the assumption that all such factors are either directly controlled for in the estimation or effectively proxied for by the lagged achievement score, which is the maintained assumption underlying value-added models. In Sec. 4.4, I will document how this channel of potential bias does not seem to be important in the case of Peru and Vietnam, where I am able to generate alternative IV estimates.

I present the estimated production function estimates for an additional grade completed in each country in Table 7. The key result is that the learning increment per additional grade completed is much larger in the Vietnamese sample than in the Indian or Peruvian samples, a conclusion that is unchanged whether or not time use categories are included. These differences are statistically significant and the learning increment per year in Vietnam is significantly greater than the increment in any of the other countries. The

difference with Peru is particularly striking since these two samples had similar baseline achievement levels at 5.

Does incorporating the differential effectiveness of schooling in the four country samples enable us to better explain the learning divergence between countries? The important pattern to note is that the inclusion of grades completed has removed the higher maturation effect (coefficient on age) in Vietnam in comparison to other countries: the pattern noted earlier, that Vietnamese children seemed to be learning more than in the other countries, disappears upon including grades completed and allowing for differential effectiveness. The contribution of this differential effectiveness of grades completed to the divergence in test scores is large: for example, raising the effectiveness of a grade of schooling to Vietnamese levels, even keeping all endowments (including learning at 5) as well as all other coefficients unchanged, closes the gap between India and Vietnam by about 70% and between Peru and Vietnam almost entirely.[22]

4.4 Are VA estimates reliable? Comparison with IV estimates

Value-added models are based on the identifying assumption that the lagged test score suffices to proxy for any relevant sources of bias in the interpretation of input coefficients in the production function estimates. Such concerns, relating to possible endogeneity of inputs, are particularly salient for grades completed in school. Within-country variation in this variable comes from three possible sources: the age of starting school, retention in particular grades due to lack of academic progress, and early or intermittent dropping out. The importance of these sources differs across the educational trajectory: whereas differences in the age of starting school account for the bulk of the variation at younger ages, in later years grade repetition and drop-out (both of which may plausibly be caused by low academic achievement) are both more relevant. If the factors that determine these three channels are effectively proxied by lagged achievement, the estimates can be interpreted causally but not otherwise. In this section, I estimate causal impacts of a grade of schooling based on a fuzzy regression discontinuity/IV approach and assess if any conclusions are substantively changed.

My strategy for estimating causal effects of additional grades completed uses variation in when children joined school, arising from their month of birth and the enrolment guidelines of particular countries, to instrument grades completed in the VA specifications presented in Table 7. Figure 4 presents the average number of grades completed by children born in different months in the sample in each country. As can be seen there is a discontinuity

[22]To see this, note that raising Indian productivity to Vietnamese levels would lead to an additional about 30 points per grade (the difference between the coefficient on grade in Col. 8 and Col. 6) which when multiplied by the mean number of grades completed (1.63 years, Table 1) closes the gap by about 49 points; this is just above 70% of the mean gap between the two countries of 70 points. The Ethiopian case is somewhat different from Peru and India since both the level and productivity of grades completed are lower than Vietnam. Here also, if the sample had the same amount of schooling and grade productivity as Vietnam, it would close about 60% of the gap between the two countries.

in Vietnam between Dec 2001-Jan 2002 and a somewhat fuzzier discontinuity in Peru between Jul-Aug 2001; these also represent the official guidelines for enrolment of children into first grade in these two countries (highlighted by red reference lines).[23]

The instrumentation strategy implies a first-stage equation of the form:

$$grades_{i,2009} = \mu + \gamma_1.Threshold_i + \gamma_2.X_i + \gamma_3.site_i + \epsilon_{i,2009} \tag{7}$$

where Threshold is defined as an indicator variable equalling 1 if born after July 2001 in Peru or after Dec 2001 in Vietnam and 0 otherwise. X is the vector of controls listed in previous specifications and includes the child's age in months. The second stage, instrumenting grades completed with $Threshold_i$, is given by the following equation which is identical to Eqs. (5) and (6) but for additionally including a vector of site dummy variables ($site_i$)within country in order to absorb any differences across sites in the implementation of enrolment guidelines.[24]

$$\begin{aligned} Y_{ic,a} &= \alpha_c + \beta_1.Y_{ic,a-1} + \beta_2.X_{ic} + \beta_3.grade_{ica} + \gamma.site_i + \epsilon_{ica} && (8) \\ &\quad +\beta_4.TU_{ic,a} && (9) \end{aligned}$$

Results from the estimation are presented in Table 8. As can be seen, in both Peru and Vietnam, the coefficients on grades completed are similar to, if somewhat smaller than, the coefficients obtained from the OLS VA models in Table 7 and coefficients on most other variables are also unchanged.[25] In both countries, the OLS VA coefficients

[23]Guidelines for enrolment in first grade in Peru in the 2007 academic year state that children should have completed 6 years of age by July 31, 2007, thus generating the discontinuity. In Vietnam, guidelines stipulate that the child should be enrolled in school in the calendar year that he/she turns six years of age, thus generating a discontinuity in grades completed between children born in December and January.

While there are similar guidelines in India as well, requiring in Andhra Pradesh all children to have turned 5 by Sept. 1 of the year in which admission is sought, the discontinuity created is much less sharp and seems inadequate in statistical power to be used as an IV by this point of the children's trajectory. Using this discontinuity, I obtain very imprecise estimates, which are not statistically distinguishable from the OLS (VA) estimates, from zero, or from the coefficients of any of the other countries, thus not allowing for any firm conclusions to be drawn; a similar conclusion is also borne by using Dec 01-Jan 02 as the relevant threshold, as used by Singh et al. (2014). Results are available on request. There is no evidence of such discontinuities that can be used in Ethiopia.

[24]The inclusion of site fixed effects is appropriate in this setting since we are not interested in comparing the constant terms across countries (unlike in previous subsections). It is useful to note, however, that it does not notably alter our conclusions even if site fixed effects are excluded from the IV specifications: the core result, of the coefficient in Vietnam being considerably higher than in Peru, is unchanged. One important difference is that in the absence of site fixed effects, the coefficient on a year of schooling in Peru is no longer statistically distinguishable from zero.

[25]In interpreting the IV estimates, it should be remembered that these are Local Average Treatment Effects identified over the compliers who are prompted to join school as a result of the discontinuity. If grade effectiveness is heterogeneous, with the youngest children in class gaining less than their older peers, a decline in the coefficient does not necessarily indicate bias. This is, however, a point of marginal concern in this particular instance since the OLS VA estimates and the IV estimates are not statistically different and the pattern across countries is entirely unchanged.

lie within the 95% confidence intervals of the IV estimates. More pertinently from our perspective, the differences indicated between the productivity of a year of schooling in Peru and Vietnam are also unchanged. In short, the VA models do not appear to be biased in these two samples.[26]

4.5 Robustness checks

Measurement error in lagged achievement

Test scores are noisy measures of (latent) academic knowledge. This measurement error in lagged achievement can cause bias in the estimated production function parameters. In order to test for this possibility, I instrumented the lagged quantitative achievement measure (CDA at age 5) with the scores of the child in a test of receptive vocabulary that was taken at the same time.[27] Informativeness of the IV rests on the correlation between different domains of cognitive achievement and first-stage results are strong. The validity of the IV rests on the assumption that the measurement error in the two tests, conducted at the same time, is independent of each other.[28]

Results are reported in Appendix B. In this sample, coefficients on the grade of schooling and other inputs are not materially affected. Coefficients on the lagged achievement measure rise in Ethiopia and Peru, consistent with attenuation bias due to measurement error in lagged achievement. The substantive results regarding the differential effectiveness of a grade of schooling in different countries and its contribution to divergence in achievement are not affected.

Flexible lag structure

In the analysis thus far, dynamics have been modelled linearly with the lagged achievement measure entering the regression specifications in levels. If growth trajectories of achievement are in fact non-linear, value-added estimates may suffer from a misspecification bias. In order to test for this possibility, I re-estimated the production function using a third-order polynomial of the lagged test score instead of the lag in levels.

[26] While this is not direct evidence supporting the validity of VA estimates in Ethiopia and India, these results are suggestive that the VA estimates are reliable.

[27] Since the vocabulary tests are administered in different languages, with corresponding differences in difficulty, I cannot directly compare them across countries. However, I can use them as instruments, utilizing the (within-country) correlation between math and vocabulary scores. Note that this instrumentation also helps deal with the issue that the test scores used in this paper are generated using an IRT measurement model and therefore include measurement error from the maximum likelihood estimation process.

[28] This is a strong assumption which rules out correlated shocks between different test outcomes, for example measurement error due to testing conditions on the day of assessment, but is often used in this literature to correct for measurement error.

Results from this exercise are reported in Appendix C. As is evident, parameters on coefficients on background covariates, time use and grades completed are not significantly altered, indicating that our key results are unchanged.

Overall, these robustness checks indicate that while there may be uncertainty regarding the 'true' value of the persistence parameter due to measurement error in the lagged measure, and the functional form in which it enters the estimation, it does not seem to change the main conclusions that are drawn in the paper from the various specifications: there is substantial divergence in learning between 5-8 years of age, even conditional on initial test scores, and it is explained by differential grade productivity in primary school and not, for the most part, by differences in the levels of inputs.[29]

5 Discussion

In this paper, I have characterized the emergence and evolution of test score gaps in quantitative ability across using comparable panel data on children in Ethiopia, India, Peru and Vietnam. Furthermore, I have decomposed the divergence of test scores between 5-8 years of age into various proximate sources.

Several results stand out. Gaps between countries open up early and show evidence of increasing substantially in the first years of schooling. Estimates from value-added models indicate that this divergence is not wholly (or even mostly) accounted for by differences in child-specific individual and home endowments, although results suggest that differences in early investments (embodied in test scores prior to school entry) have long-lasting effects. There is a significant difference in the effectiveness of schooling in the four samples which accounts for most of the gap, especially between Vietnam and the other countries. Results from VAMs seem unbiased based on comparison with IV estimates for primary school children in Peru and Vietnam.

These results have wide-ranging relevance for policy. The finding that the ranking of the four country samples in this paper is evident even at the age of 5 years, before children have begun schooling, supplements a much broader literature across disciplines in providing suggestive support for an increased focus on preschool interventions and early childhood interventions (see e.g. Grantham-McGregor et al., 2007).

However, equally importantly, I document that the gaps magnify over time with divergence in learning levels at primary school ages, especially between Vietnam and the other countries, largely being explained by the differential productivity of schooling. This indicates that there may be considerable room for corrective policy measures aimed at narrowing learning gaps in primary schools and provides a clear link between this paper

[29] In the case of Peru and Vietnam, where alternative RD/IV estimates are available, this was already shown to not be a problem since the discontinuity-based identification of grade productivity does not rely on these assumptions. These robustness checks indicate that this is also not an issue in the Indian and Ethiopian samples.

and the vast literature on such interventions in developing countries (Glewwe and Kremer, 2006; Kremer et al., 2013; McEwan, 2013); while I document that school productivity is an important source of divergence and thereby imply the need for educational interventions, the impact evaluation literature rigorously identifies the tools by which learning gains per year may be enhanced in a variety of contexts.

These results highlight the need for a shift towards emphasizing learning outcomes in developing countries rather than merely enrolment or grades completed; learning levels in these countries are low, and there are important differences in school productivity between different countries, which may offer margins for policy improvement.[30] The results also emphasize that the ranking of countries by national income (GNI) per capita and by learning outcomes are not necessarily identical: whereas Vietnamese GNI per capita (PPP $) is at a similar level as India, and less than half that of Peru, learning outcomes in Vietnam are consistently better than any of the other countries.

The analysis attempts also to show how linked panel data across countries could greatly aid the understanding of learning gaps across countries. Even though international testing programs like PISA and TIMSS are steadily increasing their coverage to also cover developing countries, as I show much of the divergence in test scores happens before the points in the educational trajectories of children where they are tested by international assessments; comparable child-level panel data could substantially complement the findings of these large representative international assessments and also guard against selection on enrolment and attendance in the estimates which is likely to be an important concern in developing countries. This may also provide a more robust basis for the comparison of learning quality across countries than estimates based on the earnings of migrants in the US as problems of differential selection across countries of origin are likely to make individual country estimates unreliable.[31]

Finally, the stark differences in the productivity of a school year across countries raise a very important question: 'Why is learning-productivity-per-year so much greater in some countries than others?' Most current work within the economics of education in developing countries focuses on the effect of particular interventions (e.g. provision of textbooks)

[30] For a detailed discussion of why policy needs to increasingly focus on learning goals rather than enrolment or inputs, see Pritchett (2013). This shift in priorities is increasingly embodied in recent policy discussions (see e.g. Muralidharan, 2013 for India) including discussions on the formulation of international development targets after the expiration of the Millennium Development Goals in 2015 (UN, 2013). This change in focus is especially relevant now given that enrolment is high in most countries and a large body of evidence has emerged that improvements in school inputs, as opposed to pedagogy or teaching reforms, have a very weak relationship to learning improvements (see e.g. Glewwe et al. 2013; Kremer et al. 2013; Das et al. 2013; McEwan 2013). That looking at quantity of schooling alone may be misleading was emphasized in an early contribution by Behrman and Birdsall (1983).

[31] For example, the estimates of schooling quality reported in Schoellman (2012) suggest that India has substantially better education than Vietnam. This is contrary both to test scores on international assessments and to findings in this paper. Plausibly, this could reflect differential selection of migrants given that Indian migrants to the US tend to be highly skilled whereas a large number of Vietnamese refugees (who were not selected on realized human capital) were settled into the US in the aftermath of the Vietnam war.

within specific contexts. While this is most useful in allowing for robust identification of policy levers that are available to national governments, it is not adequate for assessing how learning gains in a 'business-as-usual' sense differ across contexts – yet there may be important policy lessons to be gained also from asking the latter question. Our results document the difference in the average productivity of a completed grade across countries but not the sources of this differential productivity. This is an obvious area for further investigation.

References

Andrabi, T., Das, J., Khwaja, A. I., and Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics*, 3(3):29–54.

Angrist, J. D., Pathak, P. A., and Walters, C. R. (2013). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27.

Barro, R. J. (2001). Human capital and growth. *American Economic Review*, 91(2):12–17.

Behrman, J. R. and Birdsall, N. (1983). The quality of schooling: quantity alone is misleading. *The American Economic Review*, 73(5):928–946.

Berlinski, S., Galiani, S., and Gertler, P. (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics*, 93(1):219–234.

Berlinski, S., Galiani, S., and Manacorda, M. (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of Public Economics*, 92(5):1416–1440.

Bloom, N., Lemos, R., Sadun, R., and Van Reenen, J. (forthcoming). Does management matter in schools? *The Economic Journal*, forthcoming.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408.

Bloom, N. and Van Reenen, J. (2010). Why do management practices differ across firms and countries? *The Journal of Economic Perspectives*, 24(1):203–224.

Bond, T. N. and Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, 95(5):1468–1479.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632.

Cueto, S. and Leon, J. (2013). Psychometric characteristics of cognitive development and achievement instruments in Round 3 of Young Lives. Young Lives Technical Note 25, Young Lives, University of Oxford.

Cueto, S., Leon, J., Guerrero, G., and Muñoz, I. (2009). Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives. Young Lives Technical Note 15, Young Lives, University of Oxford.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., and Sundararaman, V. (2013). School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics*, 5(2):29–57.

Das, J. and Zajonc, T. (2010). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics*, 92(2):175–187.

Deming, D. J., Hastings, J. S., Kane, T. J., and Staiger, D. O. (2014). School Choice, School Quality, and Postsecondary Attainment. *American Economic Review*, 104(3):991–1013.

Engle, P. L., Black, M. M., Behrman, J. R., Cabral de Mello, M., Gertler, P. J., Kapiriri, L., Martorell, R., and Young, M. E. (2007). Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. *The Lancet*, 369(9557):229–242.

Erosa, A., Koreshkova, T., and Restuccia, D. (2010). How important is human capital? a quantitative theory assessment of world income inequality. *The Review of Economic Studies*, 77(4):1421–1449.

Escobal, J. and Flores, E. (2008). An assessment of the Young Lives sampling approach in Peru. Young Lives Technical Note 3, Young Lives, University of Oxford.

Fiorini, M. and Keane, M. P. (2014). How the allocation of children's time affects cognitive and non-cognitive development. *Journal of Labor Economics*, 32(4):787–836.

Fryer, Roland G., J. and Levitt, S. D. (2013). Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2):981–1005.

Fryer, R. G. and Levitt, S. D. (2004). Understanding the Black-White test score gap in the first two years of school. *Review of Economics and Statistics*, 86(2):447–464.

Fryer, R. G. and Levitt, S. D. (2006). The Black-White test score gap through third grade. *American Law and Economics Review*, 8(2):249–281.

Fryer, R. G. and Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, 2(2):210–40.

Glewwe, P., Hanushek, E. A., Humpage, S., and Ravina, R. (2013). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In Glewwe, P., editor, *Education Policy in Developing Countries*. University of Chicago Press.

Glewwe, P., Jacoby, H. G., and King, E. M. (2001). Early childhood nutrition and academic achievement: a longitudinal analysis. *Journal of Public Economics*, 81(3):345–368.

Glewwe, P. and Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, 2:945–1017.

Gollin, D., Lagakos, D., and Waugh, M. (2014). The agricultural productivity gap. *The Quarterly Journal of Economics*, 129(2):939–993.

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., and Strupp, B. (2007). Developmental potential in the first 5 years for children in developing countries. *The Lancet*, 369(9555):60–70.

Guarino, C., Reckase, M. D., and Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 10(1):117–156.

Hanushek, E. A. and Kimko, D. D. (2000). Schooling, Labor-force Quality, and the Growth of Nations. *The American Economic Review*, 90(5):pp. 1184–1208.

Hanushek, E. A. and Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–668.

Hanushek, E. A. and Woessmann, L. (2012a). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, 17(4):267–321.

Hanushek, E. A. and Woessmann, L. (2012b). Schooling, educational achievement, and the Latin American growth puzzle. *Journal of Development Economics*, 99(2):497–512.

Heckman, J. J. and Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6:689–733.

Hendricks, L. (2002). How important is human capital for development? evidence from immigrant earnings. *American Economic Review*, 92(1):198–219.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing TFP in China and India. *The Quarterly Journal of Economics*, 124(4):1403–1448.

Kaarsen, N. (2014). Cross-country differences in the quality of schooling. *Journal of Development Economics*, 107:215–224.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013). Have we identified effective teachers? Validating Measures of Effective Teaching using Random Assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kane, T. J. and Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Papers 14607, National Bureau of Economic Research, Inc, Cambridge, MA.

Kremer, M., Brannen, C., and Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, 340(6130):297–300.

Kumra, N. (2008). An Assessment of the Young Lives sampling approach in Andhra Pradesh, India. Young Lives Technical Note 2 2, Young Lives Project, University of Oxford.

Maccini, S. and Yang, D. (2009). Under the weather: Health, schooling, and economic consequences of early-life rainfall. *American Economic Review*, 99(3):1006–1026.

Mankiw, N. G., Romer, D., and Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, 107(2):407–437.

Manuelli, R. E. and Seshadri, A. (2014). Human capital and the wealth of nations. *The American Economic Review*, 104(9):2736–2762.

McEwan, P. J. (2013). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *unpublished*.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., and Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. ERIC.

Muralidharan, K. (2013). Priorities for Primary Education Policy in India's 12 th five-year plan. In *NCAER-Brookings India Policy Forum*, volume 2013.

Nguyen, N. (2008). An assessment of the Young Lives sampling approach in Vietnam. Young Lives Technical Note 4, Young Lives Project, University of Oxford.

OECD (2013). *PISA 2012 Results: What Students Know and Can Do - Student Performance in Mathematics, reading and Science (Volume I)*. OECD Publishing.

Outes-Leon, I. and Sanchez, A. (2008). An assessment of the Young Lives sampling approach in Ethiopia. Young Lives Technical Note 1, Young Lives Project, University of Oxford.

Pritchett, L. (2013). *The Rebirth of Education: Schooling ain't Learning*. Brookings Institution Press for Center for Global Development, Washington, D.C.

Rolleston, C. (2014). Learning profiles and the skills gap in four developing countries: a comparative analysis of schooling and skills development. *Oxford Review of Education*, 40(1):132–150.

Rolleston, C., James, Z., and Aurino, E. (2013). Exploring the effect of educational opportunity and inequality on learning outcomes in Ethiopia, Peru, India, and Vietnam. *Background Paper for the UNESCO Global Monitoring Report*.

Schady, N., Behrman, J., Araujo, M. C., Azuero, R., Bernal, R., Bravo, D., Lopez-Boo, F., Macours, K., Marshall, D., Paxson, C., and Vakis, R. (forthcoming). Wealth gradients in early childhood cognitive development in five Latin American countries. *The Journal of Human Resources*, forthcoming.

Schoellman, T. (2012). Education quality and development accounting. *The Review of Economic Studies*, 79(1):388–417.

Singh, A. (2014). Emergence and evolution of learning gaps across countries: Linked panel evidence from Ethiopia, India, Peru and Vietnam. CSAE Working Paper 2014-28, Centre for the Study of African Economies, University of Oxford, Oxford.

Singh, A. (2015). Private school effects in urban and rural india: Panel estimates at primary and secondary school ages. *Journal of Development Economics*, 113:16–32.

Singh, A., Park, A., and Dercon, S. (2014). School meals as a Safety Net: An evaluation of the Midday Meal Scheme in India. *Economic Development and Cultural Change*, 62(2):275–306.

Spencer, B. D. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement*, 20(4):pp. 317–333.

Todd, P. E. and Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, 113(485):F3–F33.

Todd, P. E. and Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, 1(1):91–136.

UN (2013). A new global partnership:eradicate poverty and transform economies through sustainable development. The Report of the High-level Panel of Eminent Persons on the Post-2015 Development Agenda, United Nations, New York.

Van der Linden, W. J. and Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In Van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of Modern Item Response Theory*, pages 1–28. Springer Verlag.

Walker, M. (2011). Pisa 2009 plus results: Performance of 15-year-olds in reading, mathematics and science for 10 additional participants. Technical report, ACER Press, Melbourne.

Figure 1: Age of children in Young Lives

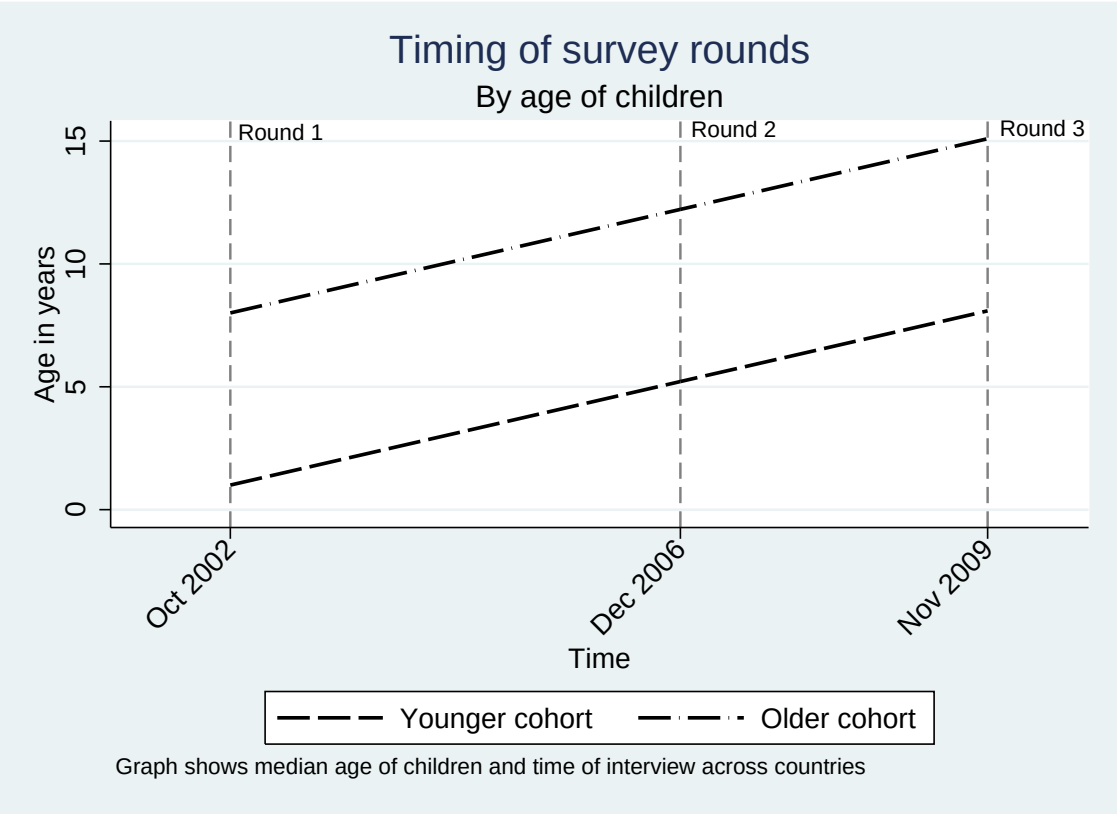
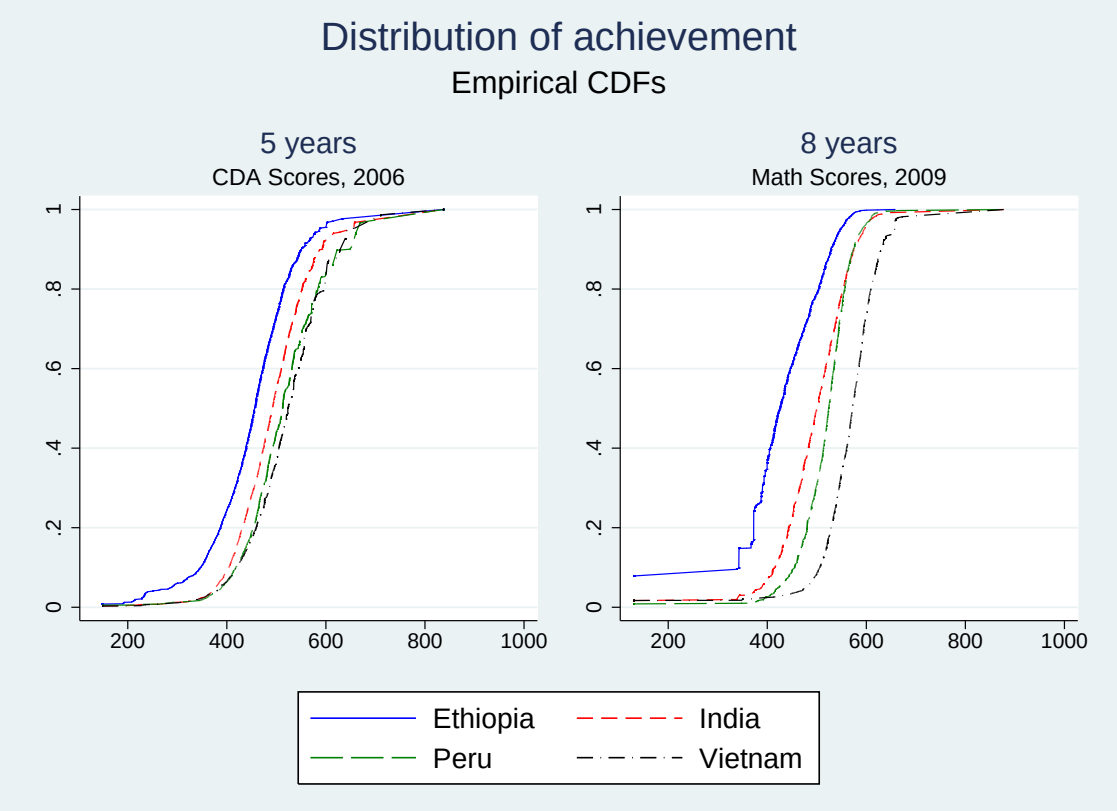
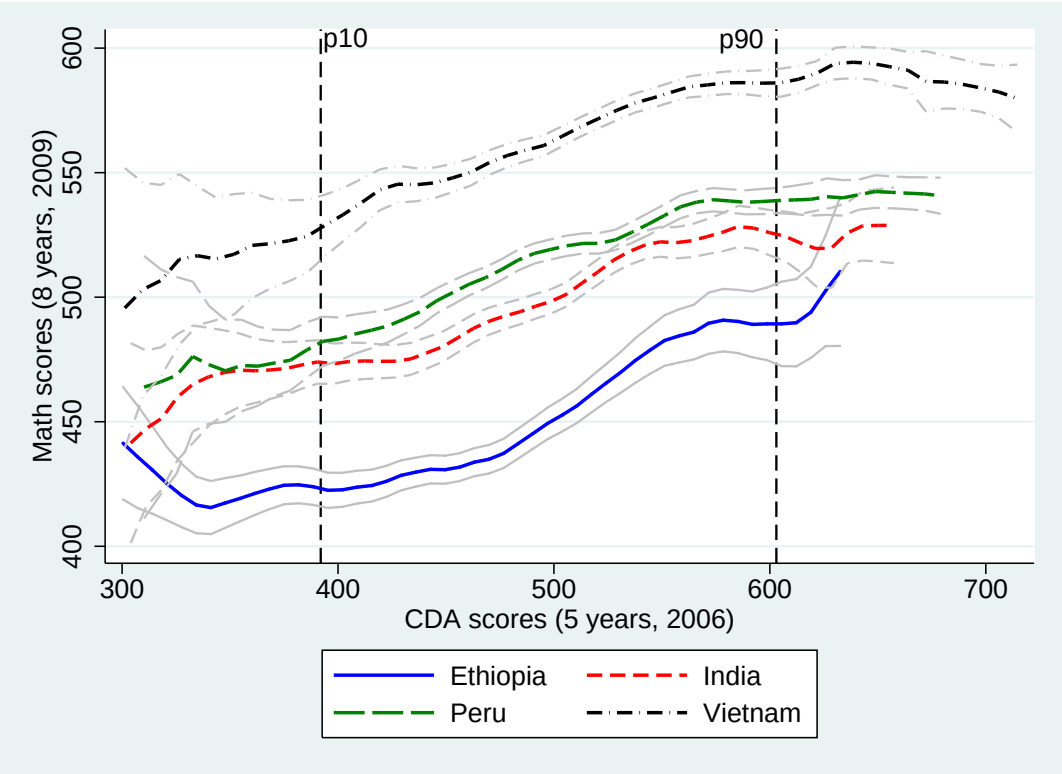


Figure 2: Learning distributions at different ages



Note: CDFs show distribution of test scores estimated with Item Response models pooling all country samples in each age group. Scores are internally normalized to have a mean of 500 and standard deviation of 100 in each age sample.

Figure 3: Progress in learning across countries: 5-8 years



Note: Lines are local polynomial smoothed lines shown with 95% confidence intervals (bandwdth of 12 points) and reference lines for relevant quantiles.The sample is restricted to observations not suffering from ceiling or floor effects in either round.

Figure 4: Discontinuity in grade attained by month of birth

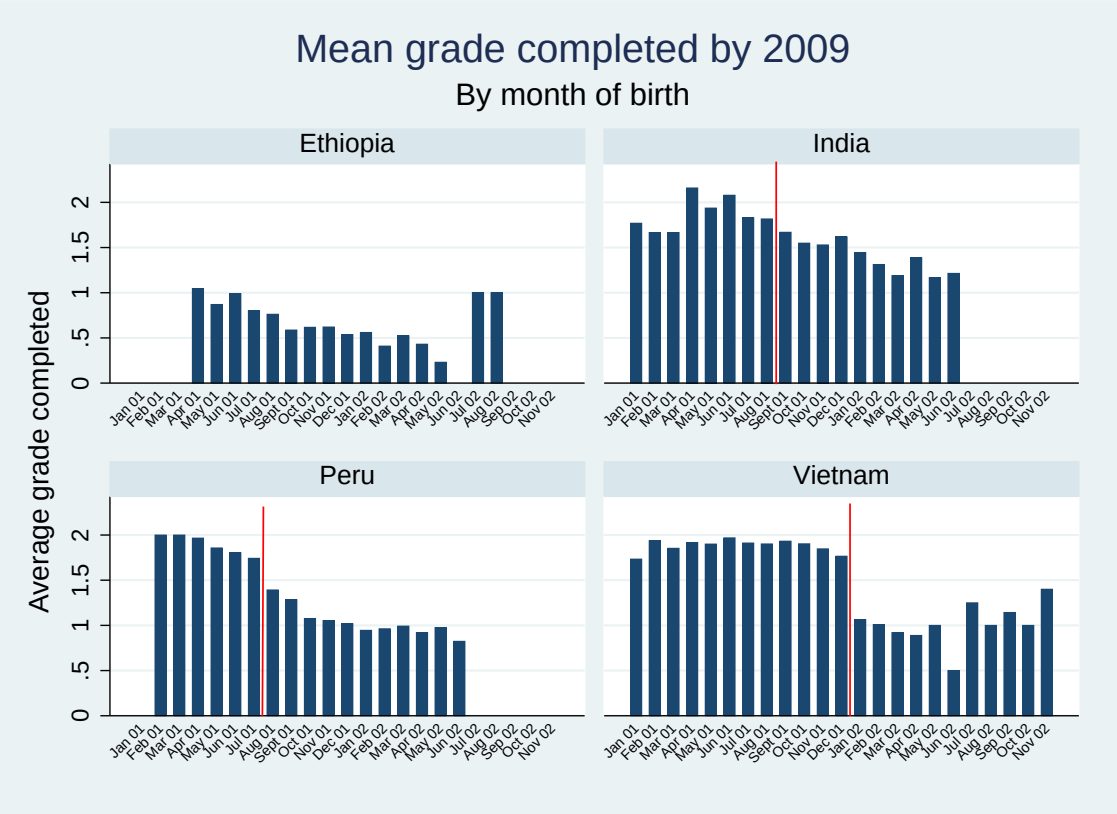


Table 1: Enrolment and grade progression of children in Young Lives

Age	Variable	Statistics	Ethiopia	India	Peru	Vietnam
5 years (2006)	Enrolment	Mean	0.04	0.45	0.01	0.01
		SD	0.19	0.5	0.1	0.08
8 years (2009)	Enrolment	Mean	0.77	0.99	0.99	0.98
		SD	0.42	0.09	0.11	0.12
8 years (2009)	Highest grade completed	Mean	0.65	1.63	1.31	1.71
		SD	0.77	1.0	0.58	0.58
8 years (2009)	Age of starting school	Mean	6.9	5.9	6.04	6.04
		SD	0.85	0.9	0.41	0.26

Note: Age of starting school is summarized only over those individuals who have enrolled in school at some point before the survey round in 2009.

Table 2: Linked test scores at 5 and 8 years

Age group	Statistics	Countries				Pooled sample
		Ethiopia	India	Peru	Vietnam	
5 years	Mean	454.1	498.8	520.9	524.5	500
	p50	456.1	491.9	512.1	523.2	495.9
	SD	102.3	94.7	97.8	89.2	100
	N	1845	1901	1893	1934	7573
8 years	Mean	417.8	495.7	516.5	566.5	500
	p50	425.8	499.6	523.9	571.1	516.7
	SD	103.7	82.7	61.9	85.3	100
	N	1845	1901	1893	1934	7573

Note: Scores are IRT test scores generated within an age sample, pooling data from all countries, and normalized to have a mean of 500 and an SD of 100 in the pooled sample. Scores are comparable across countries but not across age groups.

Table 3: Descriptive statistics of control variables

	Ethiopia			**India**			**Peru**			**Vietnam**		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Child and background characteristics (X_{ic})												
Male	0.53	0.5	1845	0.53	0.5	1901	0.5	0.5	1893	0.51	0.5	1934
First born	0.23	0.42	1845	0.39	0.49	1901	0.37	0.48	1893	0.46	0.5	1934
Caregiver's Education	2.97	3.74	1838	3.7	4.44	1898	7.79	4.62	1890	6.78	3.89	1926
Age in months	97.49	4.04	1843	96.04	3.92	1901	95.35	3.63	1892	97.09	3.75	1933
Height-for-age z-score	-1.2	1.04	1842	-1.45	1.03	1896	-1.15	1.04	1892	-1.1	1.07	1917
Wealth index (2006)	0.29	0.18	1845	0.46	0.2	1900	0.48	0.23	1893	0.5	0.2	1932
Time use (hours spent on a typical day; $TU_{ic,a}$)												
— Doing domestic tasks	1.65	1.36	1845	0.33	0.58	1901	0.87	0.7	1889	0.55	0.67	1917
— Tasks on family farm/business etc.	1.49	2.21	1844	0.01	0.25	1901	0.25	0.66	1888	0.12	0.62	1915
— Paid work outside household	0.01	0.29	1844	0.01	0.2	1901	0	0.08	1889	0	0.07	1915
— At school	4.94	2.52	1845	7.69	1.06	1901	6.01	0.93	1889	4.98	1.4	1916
— Studying outside school time	1	0.89	1845	1.85	1.09	1901	1.87	0.83	1888	2.79	1.51	1910
— General leisure etc.	4.41	2.37	1845	4.75	1.61	1901	4.15	1.67	1889	5.6	1.73	1916
— Caring for others	0.82	1.21	1845	0.21	0.52	1901	0.48	0.88	1888	0.25	0.68	1896

Note: Children in the sample were born in 2001-02. Caregiver's Education is defined in completed years; wealth index is an aggregate of various consumer durables and access to services at the household level. Height-for-age z-score is computed as per WHO standards. Unless indicated otherwise, the values of variables are from 2009.

Table 4: Do home factors and child-specific endowments explain divergence?

VARIABLES	(1)	(2)	(3)	(4)
	Dep var: Mathematics score (2009) 8-years old (Younger cohort)			
Country dummies				
India	77.8***	65.8***	62.9***	16.6***
	(3.07)	(2.98)	(3.03)	(3.65)
Peru	98.6***	80.6***	66.5***	49.2***
	(2.80)	(2.77)	(2.75)	(2.91)
Vietnam	149***	130***	111***	94.2***
	(3.10)	(3.12)	(3.02)	(3.52)
Background characteristics				
Male			3.86**	5.03***
			(1.72)	(1.63)
First-born child			6.68***	4.91***
			(1.71)	(1.64)
Caregiver's education level			3.05***	2.23***
			(0.25)	(0.24)
Age in months			3.01***	2.86***
			(0.23)	(0.22)
Height-for-age z-score (2009)			10.5***	7.81***
			(1.07)	(0.99)
Wealth index (2006)			76.1***	48.8***
			(5.47)	(5.19)
Time use (hours spent on a typical day)				
— doing domestic tasks				0.76
				(1.34)
— doing tasks on family farm etc.				-2.12*
				(1.24)
— doing paid work outside hh				0.43
				(8.74)
— at school				13.9***
				(1.01)
— studying outside of school time				13.5***
				(0.95)
— general leisure etc.				1.67**
				(0.83)
— caring for others				0.59
				(1.28)
Lagged test score (2006)		0.27***	0.12***	0.10***
		(0.011)	(0.0100)	(0.0096)
Constant	418***	295***	46.7**	-11.4
	(2.41)	(5.59)	(22.9)	(24.3)
Observations	7,573	7,573	7,522	7,465
R-squared	0.285	0.352	0.450	0.514
F-tests of equality of coefficients (p-value)				
India = Peru	0.00	0.00	0.02	0.00
India=Vietnam	0.00	0.00	0.00	0.00
Peru=Vietnam	0.00	0.00	0.00	0.00

Note: Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age.

Table 5: Country-specific production functions of achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep var: Mathematics score (2009)							
	Without time use				With time use			
VARIABLES	Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
Male	1.38	5.16	9.13***	-0.61	2.12	4.62	9.50***	-0.29
	(6.69)	(3.25)	(2.52)	(2.69)	(5.46)	(3.58)	(2.82)	(2.71)
Eldest	7.32	3.70	9.56***	7.82*	3.46	2.84	8.06**	8.37*
	(4.62)	(3.11)	(3.08)	(3.82)	(3.73)	(3.42)	(3.37)	(4.12)
Caregiver's education level	3.45***	2.98***	2.74***	3.80***	2.34***	2.30***	2.57***	2.38***
	(0.81)	(0.84)	(0.47)	(0.89)	(0.57)	(0.60)	(0.44)	(0.82)
Age in months	2.79***	2.01***	2.73***	4.15***	2.28***	2.00***	2.73***	4.25***
	(0.53)	(0.61)	(0.35)	(0.55)	(0.51)	(0.57)	(0.34)	(0.57)
Height-for-age (2009)	14.7***	10.1***	6.93***	10.5***	7.70***	9.07***	6.42***	7.47***
	(2.72)	(2.11)	(2.13)	(3.18)	(2.38)	(2.00)	(1.92)	(2.22)
Wealth index (2006)	178***	45.6	25.9***	93.6***	109***	22.0	25.8***	66.2***
	(28.3)	(27.1)	(8.04)	(28.0)	(18.8)	(20.1)	(8.20)	(20.6)
Time use (hours on a typical day)								
— doing domestic tasks					0.48	3.70	7.54***	-4.23
					(3.73)	(4.46)	(2.11)	(4.55)
— doing tasks on family farm etc.					0.68	-16.8***	-0.46	-25.2***
					(3.62)	(5.69)	(1.98)	(4.97)
— doing paid work outside hh					-4.89	22.6***	-5.88	15.8
					(9.04)	(6.69)	(5.23)	(9.40)
— at school					12.9***	22.9***	9.38***	5.28
					(3.64)	(2.70)	(3.05)	(4.94)
— studying outside school time					20.2***	20.3***	7.71***	4.74
					(3.83)	(5.14)	(1.54)	(3.49)
— general leisure etc.					1.29	5.47*	2.35*	-2.78
					(3.27)	(2.92)	(1.21)	(3.04)
— caring for others					2.15	1.00	2.02	-8.93
					(4.80)	(5.14)	(1.20)	(6.51)
Lagged CDA score (2006)	0.071**	0.15***	0.13***	0.12***	0.044*	0.15***	0.13***	0.090**
	(0.027)	(0.030)	(0.020)	(0.040)	(0.023)	(0.029)	(0.019)	(0.032)
Constant	67.9	205***	153***	36.8	51.4	-18.7	67.2*	45.7
	(48.5)	(60.5)	(32.4)	(56.8)	(64.8)	(75.3)	(33.8)	(63.1)
Observations	1,835	1,892	1,888	1,907	1,834	1,892	1,881	1,858
R-squared	0.255	0.176	0.281	0.309	0.374	0.280	0.311	0.353

Note: Robust standard errors in parentheses. Standard errors are clustered at site level. *** p<0.01, ** p<0.05, * p<0.1. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age.

Table 6: Predicted mean achievement levels under various counterfactual scenarios

		Younger cohort (8-years) Coefficients (β_c)							
		Without time use				**With time use**			
		Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
	Ethiopia	**417.55**	483.32	493.77	522.05	**417.51**	387.14	484.84	486.62
		(10.07)	(10.85)	(5.58)	(13.66)	**(11.06)**	(17.05)	(9.8)	(19.57)
Inputs	India	447.66	**495.6**	502.18	539.13	485.45	**495.6**	515.51	562.96
(X_{ic}; TU_{ica})		(11.78)	**(9.79)**	(5.06)	(11.17)	(10.59)	**(10.07)**	(8.14)	(15.09)
$Y_{ic,a-1}$	Peru	468.4	513.28	**516.41**	558.92	477.4	466.59	**516.42**	557.23
		(11.58)	(10.91)	**(4.73)**	(10.68)	(11.14)	(11.19)	**(5.76)**	(11.9)
	Vietnam	476.53	516.64	521.04	**566.69**	490.24	474.54	519.44	**567.91**
		(11.31)	(9.95)	(4.59)	**(9.31)**	(12.29)	(13.41)	(7.22)	**(11.65)**

Note: Cells contain linear predictions of test scores using combinations of country-specific production function parameters (β_c), as estimated in Table 5 with country-specific input levels (X_{ic} and TU_{ic}). Each row shows predicted values of mean achievement when applying, to a given country sample, different country-specific coefficients indicated in column headings. The diagonals in bold show the actual mean achievement in the country sample. Results are shown for specifications with and without time use categories. Standard errors of predictions in parentheses.

Table 7: Comparing effectiveness of a grade of schooling: 8-years old

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep var: Mathematics score (2009)							
	Without time use				With time use			
	Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
Highest grade completed	41.7***	28.0***	34.3***	62.1***	28.9***	25.9***	33.3***	56.3***
	(4.77)	(2.07)	(3.65)	(14.9)	(4.58)	(1.65)	(3.60)	(11.1)
Male	3.34	13.0***	8.91***	1.66	4.54	11.8***	9.10***	1.63
	(5.72)	(3.11)	(2.26)	(2.43)	(4.92)	(3.19)	(2.52)	(2.71)
Eldest	4.06	5.89**	8.61***	6.72**	1.67	4.71	7.24**	7.29**
	(4.16)	(2.65)	(2.95)	(3.04)	(3.79)	(3.07)	(3.15)	(3.21)
Caregiver's education level	3.86***	2.46***	2.27***	3.22***	2.80***	1.91***	2.14***	2.23***
	(0.67)	(0.72)	(0.50)	(0.81)	(0.53)	(0.50)	(0.49)	(0.73)
Age in months	1.29**	0.53	-0.070	0.18	1.33**	0.62	0.0067	0.71
	(0.54)	(0.46)	(0.31)	(1.13)	(0.57)	(0.42)	(0.31)	(0.89)
Height-for-age (2009)	9.49***	5.50**	5.34**	7.27***	5.41**	4.90**	4.93**	4.90***
	(2.70)	(2.26)	(1.95)	(1.81)	(2.38)	(1.89)	(1.76)	(1.59)
Wealth index (2006)	154***	54.7**	18.0*	79.8***	107***	31.7*	18.6*	60.2***
	(26.4)	(24.3)	(9.02)	(21.2)	(19.3)	(18.1)	(9.14)	(19.4)
Time use (hours on a typical day)								
— doing domestic tasks					2.28	3.11	6.88***	-4.24
					(3.44)	(4.30)	(1.99)	(3.97)
— doing tasks on family farm etc.					1.43	-13.8***	0.12	-21.8***
					(3.50)	(3.45)	(1.64)	(5.47)
— doing paid work outside hh					-3.97	22.8***	-4.34	-2.65
					(8.07)	(7.66)	(4.04)	(7.07)
— at school					12.4***	21.7***	9.10***	3.86
					(3.43)	(2.59)	(2.96)	(4.17)
— studying outside school time					14.1***	18.1***	6.86***	2.08
					(3.79)	(4.98)	(1.74)	(3.15)
— general leisure etc.					2.16	4.62*	2.43*	-2.70
					(3.30)	(2.57)	(1.34)	(2.64)
— caring for others					3.36	1.79	1.99*	-7.15
					(4.74)	(4.79)	(1.07)	(4.86)
Lagged CDA scores (2006)	0.066***	0.13***	0.10***	0.066*	0.044*	0.12***	0.100***	0.050
	(0.023)	(0.028)	(0.020)	(0.032)	(0.022)	(0.027)	(0.020)	(0.030)
Constant	188***	301***	398***	350***	120	87.9	308***	328***
	(50.3)	(46.3)	(29.8)	(75.7)	(73.5)	(54.8)	(39.3)	(66.9)
Observations	1,835	1,892	1,888	1,907	1,834	1,892	1,881	1,858
R-squared	0.340	0.276	0.343	0.437	0.410	0.365	0.369	0.458

Note: Robust standard errors in parentheses. Standard errors are clustered at site level. *** p<0.01, ** p<0.05, * p<0.1. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age.

Table 8: Discontinuity-based results on grade effectiveness in Peru and Vietnam

VARIABLES	(1)	(2)	(3)	(4)
	Dep var: Math scores (2009)			
	Peru		Vietnam	
Highest grade completed	20.6***	21.4***	48.3***	47.2***
	(7.76)	(8.12)	(7.63)	(7.29)
Male	9.62***	10.2***	1.31	1.56
	(2.44)	(2.68)	(2.39)	(2.50)
Eldest	8.20***	6.58**	5.19	6.44**
	(2.79)	(3.07)	(3.22)	(3.11)
Caregiver's education level	2.35***	2.18***	3.11***	2.46***
	(0.41)	(0.38)	(0.62)	(0.56)
Age in months	0.96	0.88	0.42	0.66
	(0.67)	(0.73)	(0.58)	(0.54)
Height-for-age (2009)	6.29***	5.73***	6.11***	4.26***
	(2.24)	(2.03)	(2.01)	(1.48)
Wealth index (2006)	30.5***	29.8***	41.1**	29.2**
	(7.80)	(7.98)	(16.6)	(13.7)
Time use (hours on a typical day)				
— doing domestic tasks		5.10***		-3.42
		(1.75)		(4.49)
— doing tasks on family farm etc.		-0.15		-15.4***
		(2.24)		(5.00)
— doing paid work outside hh		0.22		-2.42
		(4.45)		(6.41)
— at school		8.60***		7.95
		(2.94)		(5.00)
— studying outside school time		6.82***		9.72***
		(1.56)		(2.79)
— general leisure etc.		0.96		0.17
		(1.19)		(1.81)
— caring for others		2.10**		-4.48
		(0.98)		(4.15)
Lagged math scores (2006)	0.13***	0.12***	0.11***	0.089***
	(0.020)	(0.019)	(0.030)	(0.028)
Constant	285***	221***	372***	311***
	(59.5)	(70.7)	(56.6)	(61.4)
Observations	1,888	1,881	1,907	1,858
R-squared	0.365	0.393	0.480	0.504
Angrist-Pischke F-statistic	108	110	113	152

Note: Robust standard errors in parentheses. Standard errors are clustered at site level. ***p<0.01, ** p<0.05, * p<0.1. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age. Estimation includes a vector of site fixed effects, coefficients for which are not reported. Highest grade completed is treated as endogenous in this table and instrumented for using in each country a discontinuity arising from country-specific enrolment guidelines and child's month of birth.

Appendix

A Construction of Test Scores

Introduction to Item Response Theory

Test scores used in this paper are constructed using Item Response Theory (IRT) models. IRT models, used commonly in international assessments such as PISA and TIMSS, posit a relationship between a unidimensional latent ability parameter and the probability of answering a question correctly; it is assumed that the relationship is specific to the item but is constant across individuals. Further assuming local independence, conditional on ability, between answers to different items by the same person, and across persons for the same item, it is possible to write down the likelihood function for observing the full matrix of responses, given individual-specific ability parameters and item-specific characteristics; these parameters can then be recovered based on standard maximum likelihood techniques which provide unbiased estimates of individual ability.

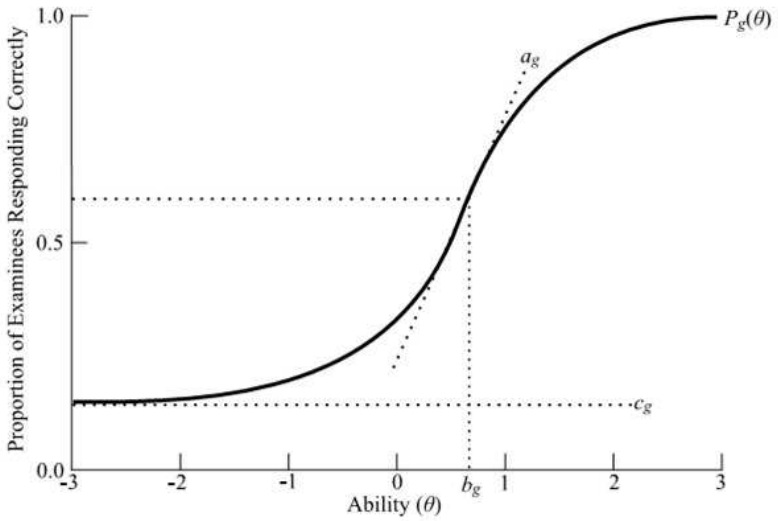
In this paper, following the procedure in TIMSS, I use the three-parameter logistic (3PL) model for all multiple-choice questions and the 2-PL model for items allowing open-ended responses. The 3-PL model is given by the following functional form:

$$P_g(X_{ig} = 1|\theta_i) = c_g + \frac{1 - c_g}{1 + exp[-1.7.a_g.(\theta_i - b_g)]} \tag{10}$$

where the probability of an individual i with ability θ_i being able to correctly answer question g is given by three item-specific parameters: the difficulty parameter b_g, the discrimination parameter a_g and the pseudo-guessing parameter c_g which accounts for the fact that with multiple choice questions even the lowest ability individual may sometimes correctly guess an answer. For the 2-PL model c_g is set to zero in which case the difficulty parameter b_g is the level of ability at which half the tested individulas would answer the question correctly.

This relationship can be depicted by plotting the relationship graphically to generate the Item Characteristic Curve, an example of which is presented in Figure A.1.

Figure A.1: Item Characteristic Curve



I used the OpenIRT suite of commands in Stata written by Tristan Zajonc to generate test scores used in this papers; specifically, I use the maximum likelihood estimates of ability for all children[32].

Testing for Differential Item Functioning

A crucial assumption underlying the use of IRT models is the absence of differential item functioning (DIF) i.e. item-specific parameters do not differ across individuals. In our application, this implies that the relationship between child ability and the probability of correctly answering a question does not differ between, say, children in Ethiopia and Vietnam. This can be a strong assumption and rules out, for example, problems due to translation of questionnaires or culture-specific framing of questions.

In order to test for the violation of the no-DIF assumption, for each item in every round of assessment, I plotted the Item Characteristic Curve based on the estimated parameters which predicts the proportion of individuals at any given ability level who will answer correctly and overlaid it with the observed proportion correct of answers at those ability levels in each country to assess if there were visible differences in Item functioning across

[32]Maximum Likelihood Estimates suffer from the problem that, while they provide unbiased estimates of the level of achievement, they overstate the variance. It is possible to use 'plausible values' estimation as used by TIMSS to generate more precise estimates of the distribution of the achievement through multiple imputation, as is done by TIMSS. However, these estimates are not unbiased estimates of individual ability and therefore cannot be used in the estimation of value-added models in the paper. For more details on Plausible Values methodology, please consult Mullis et al. (2004).

The brief explanation of IRT in this appendix draws upon Das and Zajonc (2010) and Van der Linden and Hambleton (1997). Readers should consult these sources for greater detail on IRT estimation.

country samples. For most items, there was no indication of DIF across countries; where any indication of DIF was visible, the item was 'split' in the relevant country sample i.e. treated as a separate item in the estimation of parameters and not linked to the other country samples and the IRT scores were re-estimated, following which the same procedure was repeated till no visible indications of DIF were seen. In rare cases, the probability of success did not seem to be increasing monotonically with ability (as is implied by the ICC in the estimation); these items were removed from the estimation.

Table A.1 lists the items which were split following the procedures above in each of the samples and countries. Figure A.2 presents two examples of such diagnostic graphs: as is evident, the Item in Panel A does not show any evidence of DIF whereas the Item in Panel B shows distinct evidence of DIF in India[33].

Table A.1: List of Items which were split in estimation due to DIF

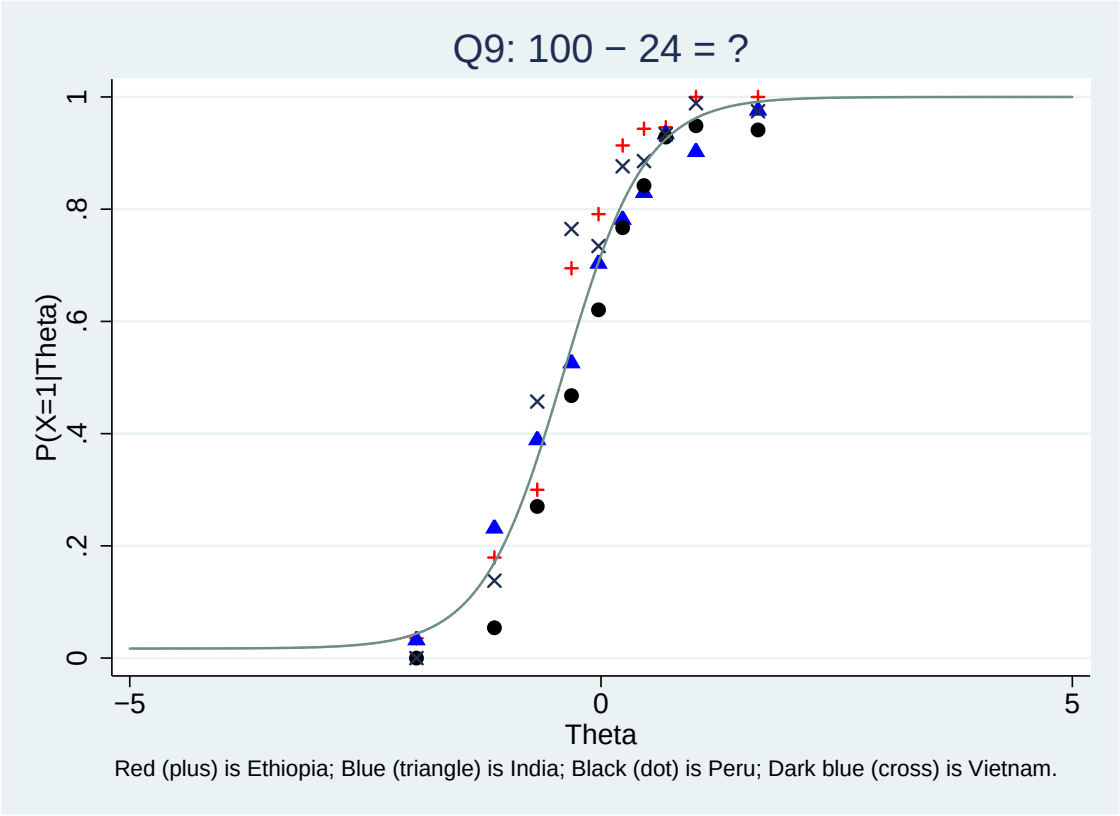
Age sample	Item no.	Countries in which modified
5-years (CDA, 15 Items)	1	Ethiopia (D), Peru (D), Vietnam (S)
	3	Vietnam (S)
	6	India (S), Peru (D), Vietnam (S)
	7	Split in all countries
	9	Vietnam (S)
8-years (Math, 29 items)	7	Vietnam (S)
	8	Peru (S), Vietnam (S)
	9	India (S), Peru (S)
	10	Peru (S)
	15	Peru (S)
	17	Vietnam (S)
	18	Peru (S)
	20	Peru (S)
	28	India (S)

(D): Item deleted
(S): Item split in estimation

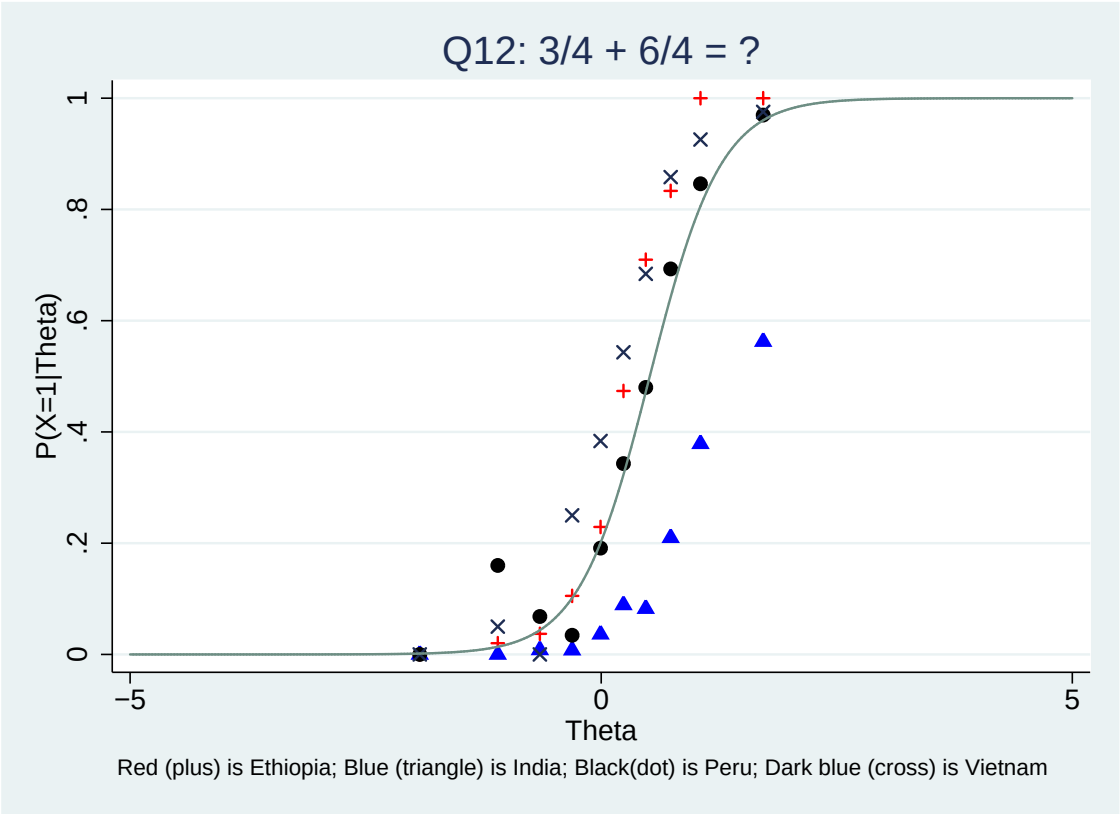
[33]Note that DIF in India also causes a poorer fit to the ICC in the other countries in panel B. This is noticeably improved after separating this question in India from the others in the estimation.

Figure A.2: Detecting Differential Item Functioning (DIF)

(a) No evidence of DIF



(b) Evidence of DIF



B Correcting for measurement error in lagged achievement

Table A.2: Estimates correcting for measurement error: 8-years old

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep var: Mathematics score (2009)							
	Without time use				With time use			
	Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
Highest grade completed	41.0***	27.4***	29.0***	63.6***	28.2***	24.9***	28.5***	58.6***
	(4.51)	(2.21)	(3.26)	(14.8)	(4.37)	(1.70)	(3.38)	(11.2)
Male	1.90	13.1***	8.34***	0.99	3.34	11.8***	8.38***	0.93
	(5.60)	(3.00)	(2.29)	(2.41)	(4.81)	(2.96)	(2.44)	(2.72)
Eldest	3.36	5.41**	7.64***	5.27**	1.03	4.71*	6.64**	6.37**
	(3.88)	(2.48)	(2.70)	(2.40)	(3.48)	(2.82)	(2.81)	(2.92)
Caregiver's education level	3.42***	2.40***	2.01***	3.08***	2.64***	1.68***	1.93***	2.25***
	(0.61)	(0.73)	(0.48)	(0.79)	(0.50)	(0.55)	(0.45)	(0.75)
Age in months	0.94	0.69	0.010	0.083	1.20*	0.68	0.048	0.64
	(0.61)	(0.48)	(0.31)	(1.20)	(0.62)	(0.44)	(0.30)	(0.96)
Height-for-age (2009)	8.96***	5.82**	3.40**	6.42***	5.31**	5.02***	3.22**	4.67***
	(2.94)	(2.36)	(1.41)	(1.48)	(2.49)	(1.92)	(1.29)	(1.50)
Wealth index (2006)	145***	57.1**	3.74	85.5***	104***	32.8*	4.97	71.3***
	(29.7)	(23.8)	(8.39)	(22.6)	(22.6)	(17.6)	(8.28)	(20.3)
Time use (hours on a typical day)								
— doing domestic tasks					2.03	3.31	6.85***	-1.97
					(3.42)	(4.32)	(1.80)	(3.20)
— doing tasks on family farm etc.					1.05	-13.7***	1.42	-19.1***
					(3.48)	(3.54)	(1.74)	(5.25)
— doing paid work outside hh					-3.71	22.7***	-5.63	-0.29
					(8.04)	(7.43)	(5.77)	(11.5)
— at school					12.2***	21.9***	7.66***	5.24
					(3.51)	(2.64)	(2.48)	(3.83)
— studying outside school time					14.2***	18.6***	6.26***	2.59
					(3.48)	(4.89)	(1.34)	(3.02)
— general leisure etc.					2.00	5.13*	2.90**	-1.19
					(3.20)	(2.68)	(1.13)	(2.60)
— caring for others					3.18	1.12	1.86*	-5.31
					(4.55)	(4.66)	(0.96)	(4.16)
Lagged CDA scores (2006)	0.15**	0.12*	0.23***	0.075	0.069	0.14**	0.22***	0.023
	(0.075)	(0.066)	(0.035)	(0.083)	(0.073)	(0.057)	(0.037)	(0.087)
Constant	188***	289***	340***	350***	128*	67.3	262***	321***
	(50.0)	(46.1)	(33.9)	(81.1)	(70.7)	(52.2)	(41.5)	(77.9)
Observations	1,821	1,821	1,848	1,708	1,820	1,821	1,842	1,662
R-squared	0.335	0.271	0.304	0.447	0.413	0.363	0.328	0.462
Angrsit-Pischke F-statistic	90.2	79.8	257	47.0	90.4	84.2	253	46.2

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Standard errors are clustered at site level. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age. Lagged CDA scores are instrumented using scores on the adapted Peabody Picture Vocabulary Test in 2006 to correct for measurement error. Coefficients should be compared to Table 7.

C Production function estimates with flexible lag specifications

Table A.3: Allowing non-linearity in lagged achievement: 8-years old

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep var: Mathematics score (2009)							
	Without time use				With time use			
	Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
Highest grade completed	41.4***	27.7***	34.0***	61.0***	28.8***	25.6***	33.1***	55.7***
	(4.85)	(2.12)	(3.77)	(14.7)	(4.67)	(1.68)	(3.70)	(11.0)
Male	2.91	12.7***	8.89***	1.41	4.22	11.5***	9.06***	1.45
	(5.59)	(3.13)	(2.33)	(2.38)	(4.92)	(3.20)	(2.57)	(2.62)
Eldest	4.21	5.31*	8.51***	6.67**	1.75	4.27	7.24**	7.37**
	(4.05)	(2.55)	(2.87)	(2.99)	(3.77)	(2.96)	(3.08)	(3.22)
Caregiver's education level	3.71***	2.41***	2.27***	3.12***	2.74***	1.85***	2.16***	2.22***
	(0.63)	(0.72)	(0.51)	(0.78)	(0.51)	(0.50)	(0.49)	(0.72)
Age in months	1.15**	0.42	-0.23	0.16	1.25**	0.53	-0.15	0.68
	(0.51)	(0.48)	(0.33)	(1.13)	(0.55)	(0.43)	(0.32)	(0.89)
Height-for-age (2009)	9.00***	5.54**	5.12**	7.10***	5.13**	4.90**	4.74**	4.69***
	(2.64)	(2.23)	(1.87)	(1.74)	(2.35)	(1.87)	(1.68)	(1.53)
Wealth index (2006)	152***	53.1**	14.5	74.8***	106***	30.4	15.5	57.2***
	(26.0)	(24.3)	(9.12)	(20.8)	(19.0)	(18.0)	(9.22)	(19.2)
Time use (hours on a typical day)								
— doing domestic tasks					2.20	3.12	7.25***	-3.81
					(3.41)	(4.24)	(2.00)	(4.11)
— doing tasks on family farm etc.					1.47	-13.8***	0.24	-21.7***
					(3.51)	(3.59)	(1.72)	(5.29)
— doing paid work outside hh					-3.83	22.6***	-5.52	-1.61
					(7.86)	(7.30)	(4.05)	(7.56)
— at school					12.2***	21.6***	9.09***	3.86
					(3.39)	(2.58)	(2.88)	(4.16)
— studying outside school time					14.1***	18.1***	6.50***	2.11
					(3.79)	(4.89)	(1.75)	(3.12)
— general leisure etc.					2.07	4.77*	2.41*	-2.46
					(3.29)	(2.54)	(1.30)	(2.61)
— caring for others					3.39	1.65	1.85	-7.04
					(4.80)	(4.81)	(1.09)	(4.90)
Lagged test score	-0.98**	-0.52*	0.20	-1.27***	-0.63*	-0.53**	0.19	-1.25***
	(0.37)	(0.26)	(0.39)	(0.33)	(0.33)	(0.20)	(0.36)	(0.36)
Lagged score, squared	0.0021**	0.0015***	0.00018	0.0028***	0.0014*	0.0014***	0.00019	0.0025***
	(0.00077)	(0.00047)	(0.00065)	(0.00064)	(0.00070)	(0.00037)	(0.00059)	(0.00072)
Lagged score, cubed	-1.34e-06**	-1.01e-06***	-3.06e-07	-1.78e-06***	-9.12e-07*	-9.55e-07***	-3.13e-07	-1.56e-06***
	(4.95e-07)	(2.69e-07)	(3.57e-07)	(4.00e-07)	(4.48e-07)	(2.29e-07)	(3.23e-07)	(4.56e-07)
Constant	357***	395***	359***	560***	226***	185***	272***	542***
	(53.1)	(67.1)	(74.8)	(88.2)	(62.1)	(63.7)	(69.6)	(92.5)
Observations	1,835	1,892	1,888	1,907	1,834	1,892	1,881	1,858
R-squared	0.344	0.280	0.353	0.442	0.411	0.368	0.379	0.463

Note: Robust standard errors in parentheses. Standard errors are clustered at site level. *** p<0.01, ** p<0.05, * p<0.1. Test scores are IRT scores normalized to have a mean of 500 and SD of 100 in the pooled four-country sample at each age. Coefficients should be compared to Table 7 which is the analogous specification entering lagged achievement linearly.