Emergence and evolution of learning gaps across countries: Linked panel evidence from Ethiopia, India, Peru and Vietnam*

Abhijeet Singh
University of Oxford

June 23, 2014

Abstract

There are substantial learning gaps across countries on standardized international assessments. In this paper, I use unique child-level panel data from Ethiopia, India, Peru and Vietnam with identical tests administered across these countries to children at 5, 8, 12 and 15 years of age to ask at what ages do gaps between different populations emerge, how they increase or decline over time, and what the proximate determinants of this divergence are.

I document that a clear pattern of stochastic dominance is evident at the age of 5 years, prior to school enrolment, with children in Vietnam at the upper end, children in Ethiopia at the lower, and with Peru and India in between. Differences between country samples grow in magnitude at later ages, preserving the country rankings noted at 5 years of age over the entire age range studied. This divergence is only partly explained by home investments and child-specific endowments in a value-added production function approach. The divergence in achievement between Vietnam and the other countries at primary school age is largely explained by the differential productivity of a year of schooling. These findings are confirmed also using an IV approach, using discontinuities in grade completion arising between children born in adjacent months due to country-specific enrolment guidelines.

*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, Jere Behrman, Santiago Cueto, 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 for providing invaluable comments on previous drafts.

Preliminary draft. Work in progress.Comments welcome at abhijeet.singh@economics.ox.ac.uk

1 Introduction

The performance of countries on comparable learning assessments differs dramatically, including among developing countries.[1] Recent research suggests that this differential performance in test scores directly translates into difference in growth rates and wage inequality (see Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008, 2012a,b; Kaarsen, 2014).

Results from comparative international tests, based on cross-sectional data on children in middle or secondary school, raise two important questions. First, at what age do gaps between countries emerge and how do they then evolve over the children's educational trajectories? Second, what are the sources of these gaps - can they perhaps be explained by systematic differences in home inputs, school quality or child-specific attributes such as ability or health across countries? These questions are central to identifying the domains in which interventions may be most required and at what ages. 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, 2013 for a recent review).

I address these questions using a unique child-level panel dataset from four developing countries – Ethiopia, India, Peru and Vietnam – collected by the Young Lives study on two cohorts of children between 2002 and 2009. The data are particularly suitable for this analysis: In each round, the same tests were administered to children in a particular age group across countries; the panel structure of the data allows me to analyze learning dynamics; and detailed household and background information allows for the estimation of rich production functions of achievement. Since the data were collected through home 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, especially in post-primary education. The data are also particularly well suited to this analysis since the four countries display remarkably different levels of achievement.

[1] For example, math scores in the 2012 PISA assessment differed by 1.4 standard deviations (SD) between Vietnam and Peru, two of the countries covered in this paper. In comparison, the difference between the US and Finland, the highest scoring Scandinavian country, was 0.38 SD while the gender gap in math performance was about 0.12 SD in the UK and Germany (OECD, 2013).

Using these data, I generate comparable distributions of test scores for quantitative proficiency across countries at the age of 5, 8, 12 and 15 years; I further link the test scores of 12-year old children to the international test distribution in mathematics from the Trends in International Mathematics and Sciences Study (TIMSS) 2003 round which covered Grade 4 students in 29 countries. I compare the achievement distributions for different country samples in the Young Lives data at each age and assess if there is a clear ranking across the samples, the magnitude of the gaps between countries, and the stability of rankings across different ages.

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. For two countries, Peru and Vietnam, I am able to use discontinuities in the number of grades completed arising from enrolment thresholds for the identification of grade productivity effects and compare results to value-added estimates; I find little evidence of systematic bias in the production function estimates compared to instrumental variable estimates based on this discontinuity.

This paper makes two key contributions. Foremost, it presents the first analysis of the emergence and evolution of gaps in cognitive achievement across countries, from the age of 5 to 15 years, 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 in test scores (see e.g. the voluminous literture 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 or covers an equivalently long age range, even within-country, using data from developing countries. Secondly, it provides the first comparable estimates of school quality across countries using individual-level panel data combined with quasi-experimental variation.[2]

[2] A small body of previous work has attempted to estimate the differential quality of schooling across countries. These papers have however relied on either the returns to schooling of immigrants from different countries to the US or on cross-sections of academic achievement from international testing programs(Hanushek and Kimko, 2000; Hendricks, 2002; Schoellman, 2012) thus precluding them from studying learning dynamics while in school. Kaarsen (2014) is a recent exception who utilizes information from cross-sections of students in Grade 4 and in Grade 8 in the TIMSS study to estimate the effectiveness of schooling in different countries thus measuring learning dynamics at the population level.

Results from the analysis are informative and often striking. *First*, learning levels in the sample are low relative to international norms: at the age of 12 years, I find that about half the children in Ethiopia, and about a quarter of children in Peru and India fail to reach the low achievement benchmark for fourth-grade children (aged about 10 years) in TIMSS.[3] *Second,* there is a clear stochastic dominance of quantitative proficiency evident at the age of 5 years, before most children have started school, with children in Vietnam and Peru at the upper end, Ethiopia at the lower end, and India in between. This ranking is stable at all ages tested i.e. at 5, 8, 12 and 15 years. Furthermore, there is clear evidence of the gaps increasing over the age range: conditional on test scores in 2006/7, in both cohorts the scores in 2009 are higher in Vietnam than in Peru which are in turn higher than in India and Ethiopia. *Third,* this gap is only partially explained by a rich measure of inputs in a production function approach including household wealth, parental education and the child's nutrition, daily time use, sex and birth order. *Fourth,* the estimated productivity of schooling differs importantly across countries: at primary school level, a year of schooling in Vietnam is considerably more productive in terms of quantitative skill acquisition than a year of schooling in Peru or India. Preferred IV estimates indicate a gain of about 0.4 SD per grade completed in Vietnam in comparison to 0.2 SD in Peru.

These results have wide-ranging relevance. 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). While a large body of evidence in this area is available from developed countries, with the exception of nutritional or health interventions, the literature on causal studies of preschool factors on cognitive achievement remains limited in other contexts.[4]

However, equally importantly, I document that the gaps magnify over time with divergence in learning levels at primary school ages, especially between Vietnam

Analysis presented here provides fresh insights, because it allows for studying individual level learning dynamics with detailed child-level information, because it analyzes gaps in cognitive achievement predating school enrolment, because it does not suffer from issues of selection resulting from increasing levels of drop-outs over the educational trajectory and because the Kaarsen (2014) study does not cover any of the four countries in this paper.

[3] As a comparison, only 7 percent in the UK and the US, and less than 3 percent in Singapore and Hong Kong, fail to reach this basic level in fourth grade when they are aged 10 years on average.

[4] 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.

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. Foremost, this highlights 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.[5] 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 only about a third of Peru, learning outcomes in Vietnam are consistently better than any of the other countries.

Finally, the stark differences in the productivity-per-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) 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.

The analysis and results presented here relate to several strands of the literature within economics. Methodologically, this paper is closest 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) and to the literature on value-added models of achievement using household-based panel data (see e.g. Todd and Wolpin, 2003, 2007; Fiorini and Keane, 2014). Additionally, the methodology of the paper also draws upon the literature on mapping test scores on a comparable metric using Item Response Theory (IRT) models which are commonly used in comparative educational

[5]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).

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; Andrabi et al., 2011).

The results speak directly to a large, and rapidly growing, literature that documents low levels of learning in developing countries (see Glewwe and Kremer, 2006 for an authoritative review) and experiments with different interventions to improve these low levels of learning (see Kremer et al. 2013 and McEwan 2013 for meta-analyses). Results on differences in the productivity of a school year in producing test scores relate directly to studies seeking to explain the effect of differential schooling quality to growth (Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008; Schoellman, 2012; Kaarsen, 2014) and echo similar questions from the cross-country growth literature about why productivity per worker differs vastly across countries (e.g. Hall and Jones 1999); indirectly, differences in the productivity of schooling may relate also to differences in the quality of management across different countries (Bloom and Van Reenen, 2007, 2010; Bloom et al., 2014).[6]

The rest of this paper is structured as follows: Section 2 describes the data used in this paper and puts the samples in context by comparing achievement with the international test distribution in TIMSS 2003; Section 3 investigates stochastic dominance of test outcomes across age groups and assesses whether learning gaps seems to narrow or widen over time between countries; Section 4 estimates value-added models of achievement, presenting assessments of school effectiveness and of inter-group differences in achievement; it also presents some 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 ('OC' hereafter, born

[6]In seeking to use comparable micro data to study cross-country differences in productivity, analysis in this paper also resembles similar recent attempts in other economic sectors.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.

in 1994/95) was aged about 8 years and the younger cohort ('YC' hereafter, born in 2001/2) was aged between 6-18 months at the time of the first wave of the survey in 2002. In each country, 2000 children of the younger cohort and 1000 children in the older cohort were surveyed.[7] 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.[8] 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 more than 90% of the original sample still in the survey in 2009 in each country. In this paper, I use test data on quantitative skills for both cohorts from the 2006 and 2009 rounds.[9] Figure 1 presents the age of children in the two cohorts at each survey wave[10].

Table 1 presents descriptive information about the educational trajectories and progression of children in the different age samples in Young Lives at different ages which will be of central importance in interpreting all results in this paper. Three 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; 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 an average above 7 years); and the rates of drop-out by the age of 15 also importantly vary with the highest dropouts being in India and Vietnam where just under a quarter of the 15-year old sample is no longer enrolled in school.

In each round, a range of background information and child-specific data, including cognitive and nutritional outcomes, was collected.[11] In 2006, quantitative skills were tested for the younger cohort (then aged about 5 years) using the 15-item Cognitive Developmental Assessment tool developed by the

[7]The only exception is Peru where only 716 children in the older cohort were surveyed due to resource constraints.

[8]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).

[9]The 2002 round had limited achievement data. On quantitative proficiency, only a single test item ('2 x 4 = ?') was administered.

[10]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.

[11]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.

International Evaluation Association Preprimary Project; a 10-tem Mathematics test was administered to the older cohort, then aged about 12 years, which included six math problems from the publicly released items from the fourth grade tests of the Trends in International Mathematics and Science Study (TIMSS) 2003 round. In 2009, a 29-item mathematics test was administered to the younger cohort and a 30-item test to the older cohort.[12]

The analysis in this paper requires generating comparable test scores for each cohort/round sample and further, for the 12-year old sample, generating test scores that are directly comparable to test scores from the TIMSS 4th grade sample. This is achieved using Item Response Theory (IRT) models which 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. 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. The survey instruments, including tests, were harmonized across countries in each round, allowing us to generate test scores on a comparable scale across countries.

IRT models 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 linking across samples even with only partial overlap of test items; 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.

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, we cannot compare the absolute levels of achievement across two samples. Since tests were not harmonized across rounds or across cohorts in Young Lives, I cannot link test scores on a comparable scale over time or across cohorts.

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

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. In this paper, maximum likelihood estimates of ability are used.

2.2 What are absolute learning levels in these samples?

In this section, I compare the performance of the12-year old sample in the Young Lives data to the performance of fourth grade students covered by the TIMSS assessment in 2003 who were administered a subset of the same questions. This is useful for putting these samples in context but also is valuable information in its own right: none of these countries have been covered by the TIMSS 4th Grade assessment previously although some comparisons at 15 years of age exist.[13]

Table 2 presents the proportion of test-takers in the Young Lives sample, and in the TIMSS 4th Grade sample from select countries which had participated in the 2003 round, who answered the six common questions correctly. As is clear from the Table, a smaller proportion of 12-year old children in the Young Lives sample answer the common test questions correctly than 10-year old children in most OECD economies. The only exception is Vietnam where the proportion correct across questions seems to be comparable to many OECD country samples and is significantly higher than in the other Young Lives countries for most questions.

Table 2 is inadequate to compare the full learning distribution across the Young Lives sample, or to the TIMSS sample, as this requires the aggregation of responses to different test items into a single test score. In order to allow such comparisons, I use the six common items from TIMSS as anchor items and generate comparable IRT test scores.[14]

Before presenting the IRT estimates, it is useful to point out that the assessment in this age group is affected importantly by ceiling and floor effects, thus not capturing the full spectrum of ability. About 33% of children in Vietnam, 18% in India, 8% in Peru and 4% in Ethiopia answered all ten questions correctly while about 8% of children in Ethiopia, 6% in India, 3% in Vietnam and 1% in

[13]The PISA assessment at the age of 15 years has previously covered Peru (in 2000, 2009 and 2012), Vietnam (2012) and two states in India (2009), although not Andhra Pradesh the state covered by the Young Lives study. Ethiopia has never been covered by TIMSS or PISA.

[14]Item characteristics, i.e. the item specific parameter values, were taken from the TIMSS 2003 report for the common items and treated as known with the item characteristics of the other questions and the ability of the sample individuals treated as unknown parameters to be estimated. TIMSS reports test scores by rescaling the proficiency estimates to have a mean of 500 and a standard deviation of 100 for their international sample; I follow the same procedure in order to keep test scores exactly comparable to the TIMSS sample.

Peru answered none correctly. IRT maximum likelihood estimates are undefined when respondents answer fewer questions correctly than would be expected by guessing or answer all questions correctly. The common solution adopted in practice to address this issue is to bound the upper and lower limit of the ability distribution.[15] For this sample, scores are bounded between [0,1000]. Since the mean of the distribution may be affected by these ceiling and floor effects, I have restricted myself to only comparing the performance of the median child for this age sample.[16]

Table 3 presents the median of the distribution of mathematics achievement in the 12-year old sample and the proportion of children who fall below the low and intermediate benchmarks developed by TIMSS and linked to their test scale for mathematics.[17] The median child in Ethiopia, India and Peru is below the median in the TIMSS 4th Grade sample and with significant gaps in proficiency compared to children in OECD economies.

More meaningfully, over a quarter of the sample in India and Peru, and about half in Ethiopia, fail to reach even the Low International Benchmark, defined by TIMSS as follows: *"Students have some basic mathematical knowledge. Students demonstrate an understanding of whole numbers and can do simple computations with them."* (Mullis et al., 2004). In Peru and India, the median child is behind the intermediate benchmark for fourth-graders even though he is in 6th and 7th grades respectively in the two countries; despite the 2-3 additional years of schooling, this is notably worse than for any of the selected countries from the TIMSS sample displayed in Table 3.

Table 3 and the foregoing discussion highlight two patterns quite starkly: that there exist vast differences in the mathematical knowledge acquired by children of the same age across the four countries in the Young Lives samples; and that the absolute levels of learning, with the possible exception of Vietnam, are very low in comparison to OECD countries, even though in Peru and India the children have had two extra years of schooling compared to the OECD sample.

[15] For example, TIMSS and Das and Zajonc (2010) bound the ability distribution between [5,995] and until their recent revision of scales, GRE scores were bounded between [200,800].

[16] This bounding does not affect the qualitative results presented in this section which are also borne out by plausible value estimates not subject to the same concerns. For the other age-samples presented in this paper, this bounding exercise affects only a small proportion of the sample as the tests provide a smooth measure of achievement with no ceiling or floor effects. The censoring of achievement for the 12-year old sample will also present challenges for the estimation of value-added models later since the inclusion of censored regressors can induce bias; I will test for this concern explicitly in Section 4.

[17] TIMSS defines four international benchmarks – Low (400), Intermediate (475), High (550) and Advanced (625) – each with descriptions of the level of skills. See Mullis et al. (2004) for details.

3 When do gaps emerge and how do they evolve?

The previous section documented that by the age of 12 years, there is a substantial difference in the performance of children in the four study countries on a mathematics test. In this section, I analyse whether a similar pattern is also discernible at other ages and whether there are not only systematic levels differences between children in different countries but also differences in how much they learn over time.

Table 4 presents descriptive statistics of the quantitative ability scores at each of the other three ages at which testing was done i.e. at the age of 5 years, 8 years, and 15 years. 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; test scores are not linked across cohorts or over time and therefore cannot be directly compared across age groups.

The most striking pattern in Table 4 is that the ranking of countries seen at the age of 12 years is already evident at the age of 5 years, which pre-dates schooling for most of the sample. Given the literature from OECD countries, combined with a growing literature from developing countries which also documents the cognitive effects of early childhood influences (e.g. Glewwe et al., 2001 and Maccini and Yang, 2009 on the effect of nutrition in early childhood) on educational achievement, 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) remains very limited in developing countries.

The second notable pattern is that the ranking of countries in the Young Lives sample is unchanged across age groups. Although occasionally differences between countries are not statistically significant – at the age of 5 between Vietnam and Peru, and at the age of 12 between Peru and India – 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 where Vietnamese children score significantly higher than children in Peru and Peruvian children score significantly higher than Indian children.

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. The CDFs of the estimated test

scores are plotted for each of the four age groups in Figure 2. As may be seen, conclusions formed on the basis of the mean comparisons also hold true across the entire distribution: there is a clear pattern of stochastic dominance with Vietnamese children performing better at every age compared to other samples and with Ethiopian children performing distinctly worse.[18] With the anomalous exception of the 12-year sample, the Peruvian CDF always lies to the right of the Indian CDF. In general, there seems to be a clear and stable ranking with Vietnam >Peru >India >Ethiopia in these samples.

Since test scores are not linked across time and across cohorts, Figure 2 is not enough by itself to comment on whether differences between country samples are exacerbated as children grow older. Further, even if gaps have grown, Figure 2 does not 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 Figures 3a and 3b, I present non-parametric plots of achievement in 2006 and achievement in 2009 for the four country samples in both cohorts. 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)?

Between 5 and 8 years of age (Figure 3a), there seems to be considerable divergence between countries 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. Between the age of 12 and 15 (Figure 3b), however, a somewhat different picture emerges: children in Vietnam and Peru display near-identical trajectories of achievement which are higher than children in Ethiopia and India which are almost indistinguishable from each other. The difference between the two sets of countries seems to be a difference in the intercepts and not the slopes[19].

[18] In generating Figure 3, and in all subsequent analysis in this paper, I have re-normalized the 12-year old achievement to have a mean of 500 and SD of 100 in the pooled Young Lives sample (rather than linked to the TIMSS normalization as in Section 2) in order to keep normalization procedures identical for each age group. Ceiling and floor effects in the 12-year old sample can be noted in the empirical CDFs.

[19] Similar patterns of divergence in these data are also documented in Rolleston (2014) and Rolleston et al. (2013). 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 as used by both

4 Sources of divergence: Results from value-added models

Analysis presented in the previous section documents the pattern of divergence across different country samples but 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:

$$\begin{aligned} Y_{ic,a} &= \alpha + \beta_1.\theta_c && (1)\\ &\quad +\beta_2.Y_{ic,a-1} && (2)\\ &\quad +\beta_3.X_{ic} && (3)\\ &\quad +\beta_4.TU_{ic,a} + \epsilon_{ica} && (4) \end{aligned}$$

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 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 category).

Rolleston (2014) and Rolleston et al. (2013). Despite differences in method, however, the basic descriptive findings are similar.

Rolleston (2014) and Rolleston et al. (2013) do not attempt comparisons of individual performance in the Young Lives sample with the performance of the international test distribution in TIMSS or an analysis of the sources of divergence across countries.

The estimation is carried out separately for the two cohorts (i.e. at ages 8 and 15) by pooling all country samples within cohort.

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.[20]

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). 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. Lagged value-added 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.

As Todd and Wolpin (2003, 2007) discuss, this strategy depends on strong assumptions (e.g. geometric decay of inputs) and may suffer bias from measurement error and unobserved heterogeneity. The observed level of bias in input parameters estimated by value-added models, however, seems to be low in practice across a range of applications including in comparisons with experimental estimates (Kane and Staiger, 2008; Kane et al., 2013; Deming et al., ming; Angrist et al., 2013; Singh, 2013), with quasi-experimental estimates (Chetty. et. al., forthcoming) , dynamic panel data estimates (Andrabi et al., 2011) and in simulated data with a variety of non-random assignment mechanisms (Guarino, Reckase and Wooldridge, forthcoming) . These VAMs will be used as the workhorse specifications for the analysis of divergence in

[20]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.

The bounding exercise is also important in accounting for differences in enrolment across countries at different ages: as Table 1 documents, that nearly a third of the children in the Ethiopian sample are not yet enrolled at the age of 8 years while by the age of 15 drop-out rates are higher in India and Vietnam than in the other two countries.

this paper although I will investigate possibilities of bias due to unobserved heterogeneity and measurement error later in this section.

Summary statistics of the controls used in the estimation of achievement production functions are presented in Table 5. Results from the estimation of Specifications 1-4 are presented for both cohorts in Table 6. In both cohorts, the mean differences in the test scores are statistically different across the four countries and substantial in magnitude (Cols. 1 and 5). Controlling for the lagged test achievement (Cols. 2 and 6) reduces the gap between countries somewhat in the younger cohort and substantially in the older cohort (eliminating most of the gap between Peru and Vietnam and between Ethiopia and India). Gaps decline further upon inclusion of background variables in X but the magnitude of decline is small as a proportion of the initial gap.

Inclusion of time use inputs has different effects in the two cohorts. In the older cohort aged 15, the gap between Ethiopia and Vietnam increases while in the younger cohort aged 8, 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 latter 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, and because children in Vietnam and India are more likely to have left schooling by 15 than in the other two countries.

The central pattern in Table 6 is that a substantial gap remains between the country samples; even in the most extensive specifications, Ethiopia and Vietnam differ by between 0.7-0.9 SD at 8 and 15 years of age, accounting for more than 60% of the cross-sectional gap in test scores. 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 for 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 Tables 7 and 8 for the 8 year old and 15 year old sample respectively.

As can be seen the coefficients on specific inputs differ greatly across countries. It is, however, difficult to directly read from these Tables the importance of this differing productivity for explaining test score gaps. In order to facilitate such comparison, I present some counterfactual examples applying to each country sample, the input coefficients estimated in the different country samples i.e. predicting the mean level of achievement keeping the country's level of inputs fixed but varying the coefficients of the inputs (including the constant term) to match other countries.[21] The results are shown in Table 9.

The results are informative and telling. In the younger (8 year old) sample, the difference between the average levels of achievement between Ethiopia and India 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 7 for the country, reduces the gap between Ethiopia and India by about two-thirds.[22] However, strikingly, 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 the level of inputs. For example, while raising Ethiopian inputs to Vietnamese levels only closes 45% of the observed test score gap between the countries, equalizing the productivity of the inputs closes about three-quarters; similarly, about 70% of the gap between India and Vietnam is covered by equalizing the productivity of inputs to Vietnam.

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

[21] 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 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.

[22] 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.

Vietnam and the rest comes from the difference in the coefficients on the age in months in Table 7: 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[23]; this will be investigated in the later subsections.

Results for the older cohort are rather more mixed. The gap between Ethiopia and India seems to be entirely a product of the differences in the endowments across samples including the amount learnt till the age of 12 years. In closing the gaps between Vietnam and the other countries as well, other than perhaps in India, equalizing the level of all inputs seems as effective as equalizing all production function parameters.

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

As Table 9 documents, only a small portion of the divergence across countries (especially with Vietnam) till the age of 8 years is accounted for by the levels and differential productivity of home inputs across the four country samples. 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 (Hanushek and Kimko, 2000; Schoellman, 2012; Kaarsen, 2014) 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

[23]The difference between the coefficient on age in Vietnam and in the other countries is invariably statistically significant in cross-equation tests.

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.[24]

I present the estimated production function estimates for an additional grade completed in each country in Table 10 for the younger cohort. The results for younger children are striking: the learning increment per additional grade completed is much larger in the Vietnamese sample than in the Indian sample, 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.

Does incorporating the differential effectiveness of schooling in the four country samples enable us to account for a larger proportion of the 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 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 entirely.[25]

Results for the older cohort (Table 11) are mixed and stand in contrast to the results at 8 years of age: in particular, the progress in Vietnam appears lower than in other countries[26]. This is surprising and contrasts with all the previous patterns highlighted in the data. A likely possibility which could account for

[24]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 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.

[25]Ethiopia is somewhat an exception 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.

[26]This pattern is not entirely robust. In specifications which include time use categories, the coefficient on highest grade schooling in Vietnam is not significantly different from India or Peru.

this anomalous finding is that as children move to secondary school, the focus of math training moves away from arithmetic and basic geometry (the focus of the testing in Young Lives) to more challenging topics such as trigonometry, probability and calculus which are not assessed here; if such a movement is more pronounced in Vietnam than elsewhere, or if Vietnamese children had already substantially mastered the range of skills tested by the age of 12, it is possible that we do not see much progress at all on the available test metric[27].

The important point to stress regarding the divergence between 12 to 15 years is that the differential productivity of schooling, which is the most salient contemporaneous policy variable among the inputs in the production function, does not account for the divergence between country samples.

4.4 Are VA estimates reliable? Comparison with IV estimates

As noted above, 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, by the age of 15 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 plausibly exogenous variation arising from enrolment guidelines 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

[27]Note that the coefficient in Ethiopia is invariably larger than the coefficient in the other countries both with and without time use; this difference is statistically significant from Vietnam in both specifications and from India in the specification with time use. Combined with the fact that the median grade completed for this age group in the Ethiopian sample is Grade 6 while in the Indian and Vietnamese samples it is Grade 9, this could be indicative that the skills tested by the Young Lives assessement are mostly produced at lower grade levels.

in the VA specifications presented in Table 9. 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 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).[28]

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

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

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 site fixed effects ($site_i$)within country in order to absorb any differences across sites in the implementation of enrolment guidelines.[29]

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

Results from the estimation are presented in Table 12. As can be seen, in both Peru and Vietnam, the coefficients on grades completed are similar to, if

[28]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).

There is also no evidence of such discontinuities that can be used in Ethiopia.

[29]The inclusion of site fixed effects is appropriate in this setting since we are not 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.

somewhat smaller than, the coefficients obtained from the OLS VA models in Table 9 and coefficients on most other variables are also unchanged.[30] In both countries, the OLS VA coefficients 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.[31]

4.5 Robustness checks

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, following the practice in Chetty et al. (ming).

Results from this exercise are reported in Appendix B. In the 8-year sample, parameters on coefficients on background covariates, time use and grades completed are not significantly altered, indicating that our key results are unchanged. In the 15-year old sample, there are some differences in the coefficient on grades of schooling which declines in India and Peru but the main comparative conclusions about rates of progress across countries (or the contribution of different factors to divergence between 12 and 15 years of age) are not changed.

Bias due to censoring in the lagged achievement score for 15-year olds

As noted in Section 2 and 3, there are non-trivial ceiling and floor effects in the 12-year old maths test, leading to a censoring of the true achievement

[30] 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 the effects of grade effectiveness are 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.

[31] While this is not direct evidence supporting the validity of VA estimates in Ethiopia and India, it is suggestive that the production function estimates are reliable. Unfortunately, these discontinuities seem to lack explanatory power in the first-stage at the age of 15 years. So I cannot similarly test the validity of the VA estimates for the older cohort.

distribution. Recent work by Rigobon and Stoker (2007, 2009) demonstrates that such censoring in regressors can cause bias in estimates; given that the severity of the censoring varies across the four country samples, it is possible that different degrees of bias could exist and affect the comparative results in this paper for 15-year olds.

To test the empirical relevance of this bias, I use Bayesian *expected a posteriori* (EAP) IRT scores from the 12-year old assessment instead of maximum likelihood estimates of ability as the lagged achievement measure and rerun the estimation reported in Table 11. EAP scores utilize information from a prior distribution of ability along with the log likelihood and are less affected by censoring issues.[32] The resulting estimates are presented in Appendix C. As may be noted, although the coefficients on the lagged achievement measure rise markedly in both Vietnam and India (which were most affected by censoring) most input parameters, including most importantly the coefficient on grade completed, are unchanged indicating that our conclusions continue to hold. Notably, the coefficient on lagged achievement looks considerably more similar across countries correcting for the censoring than in Tables 8 and 11.

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 in the younger cohort, 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.[33] 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.[34]

[32]Please see Das and Zajonc (2010) for an explanation of technical details of Bayesian EAP estimation of IRT models.

[33]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.

[34]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.

This exclusion restriction is not maintainable in the older cohort due to the censoring of lagged achievement. Since the censoring of lagged achievement necessarily implies that measurement error is correlated with ability, an independent measure of ability (such as the vocabulary test) cannot be a valid IV for the lagged achievement.

Results are reported in Appendix D. In this sample, coefficients on the grade of schooling and other inputs are not materially affected. Coefficients on the lagged achievement measure rise significantly in Ethiopia and Peru, consistent with attenuation bias due to measurement error in lagged achievement, but remain statistically indistinguishable from OLS results in India and Vietnam. 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.

Overall, the range of 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 divergence in both age groups even conditional on test scores and it is explained by differential grade productivity in primary school between 5-8 years of age but not in the older cohort between 12-15 years.

5 Conclusions

In this paper I have characterized the emergence and evolution of test score gaps in quantitative ability across using panel data on two cohorts of children in Ethiopia, India, Peru and Vietnam. Furthermore, I have decomposed the divergence of test scores, between 5-8 years of age and between 12-15 years of age, into various proximate sources.

Several results stand out. Achievement levels in three of the four countries are very low by international standards. Gaps between countries open up early and show evidence of increasing over the educational trajectory, thus preserving the ordering of country samples apparent at the age of 5 years. 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 significant difference in the exposure to and effectiveness of schooling in the four samples which accounts for an important portion of the gap, especially at primary level. Results from VAMs seem unbiased based on comparison with IV estimates for 8-year old children in Peru and Vietnam.

It is important that the differences in the levels of test scores at 8 years are accounted for almost entirely by the difference in the quality of schooling across countries: this matters particularly because school-based learning may be easier for policy to directly influence than (potentially unobserved) investments into children at home. This also provides a clear link between this paper 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. The persistence of gaps between the ages of 12 and 15 seems likely to be a product of investments made prior to 12 years i.e investments in early childhood and primary school.[35]

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 to the US as problems of differential selection across countries of origin are likely to make individual country estimates unreliable.[36]

Finally, it should be noted that the results tell us the difference in the average productivity of each completed grade in the different countries but not the sources of this differential productivity at the school level. This is an obvious area for further investigation.

[35] In some respects these findings are similar to structural analyses in the US but applied here to a comparative cross-country setting. See, for example, Cameron and Heckman (2001) and Keane and Wolpin (2001) who report that relaxing credit constraints at the age of 16 years does not achieve much in increasing college enrolments and that most differences in enrolment decisions seemed to predate from background factors and childhood investments.

[36] For example, the estimates of schooling quality used 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.

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.

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. (2014). Does management matter in schools? Discussion Paper 13-032, Stanford Institute for Economic Policy Research (SIEPR), Stanford,CA.

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.

Cameron, S. V. and Heckman, J. J. (2001). The Dynamics of Educational Attainment for Black, Hispanic, and White males. *Journal of Political Economy*, 109(3):pp. 455–499.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (forthcoming). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, forthcoming.

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.

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.

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., Hastings, J., Kane, T., and Staiger, D. (forthcoming). School choice, school quality and academic achievement. *The American Economic Review*, forthcoming.

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.

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

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):forthcoming.

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. (forthcoming). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, forthcoming.

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114(1):83–116.

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. (2013). The economics of human development and social mobility. *unpublished*.

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.

Keane, M. P. and Wolpin, K. I. (2001). The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review*, 42(4):1051–1103.

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*.

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.

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.

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*, 1:1–37.

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

Rigobon, R. and Stoker, T. M. (2007). Estimation with censored regressors: Basic issues. *International Economic Review*, 48(4):1441–1467.

Rigobon, R. and Stoker, T. M. (2009). Bias from censored regressors. *Journal of Business & Economic Statistics*, 27(3):340–353.

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*.

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

Singh, A. (2013). Size and sources of the Private School Premium in test scores in India. Young Lives working paper, Young Lives, University of Oxford, Oxford.

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.

Figure 1: Age of children in Young Lives

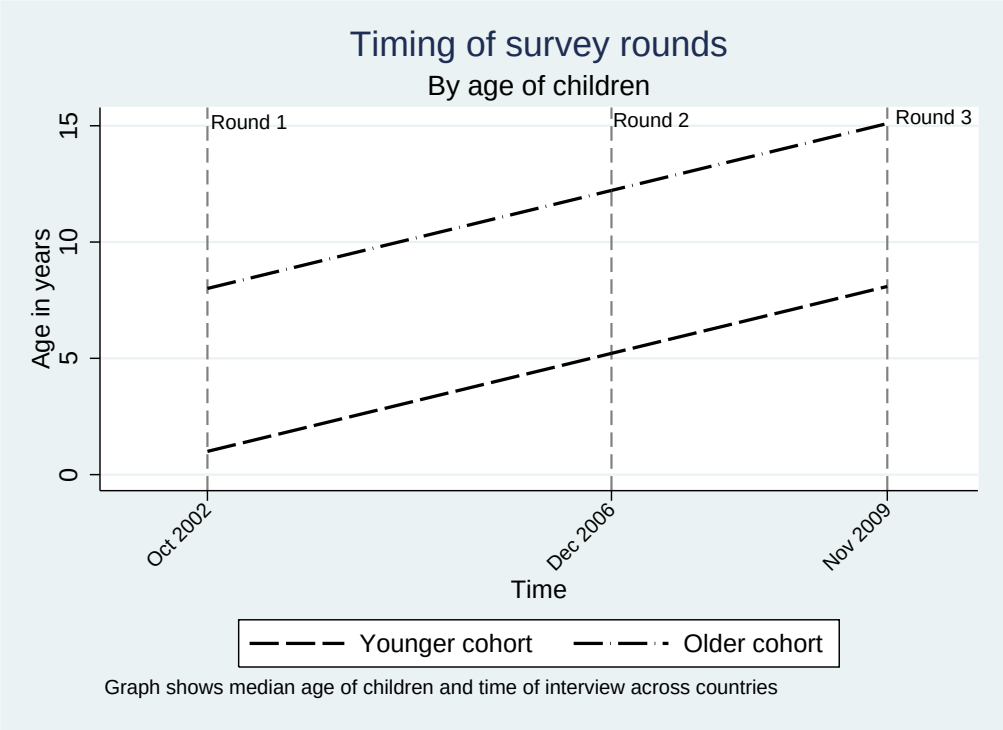
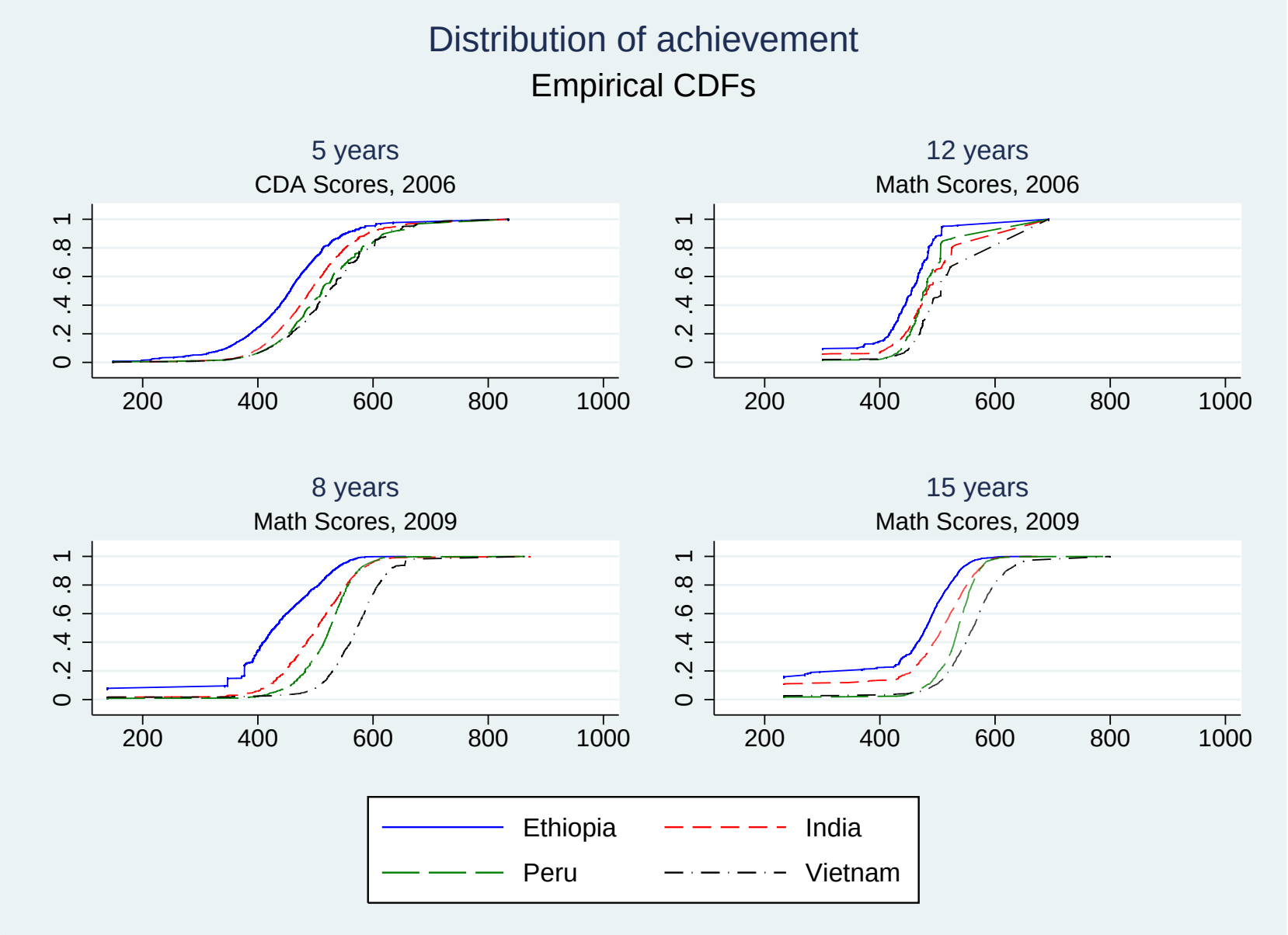


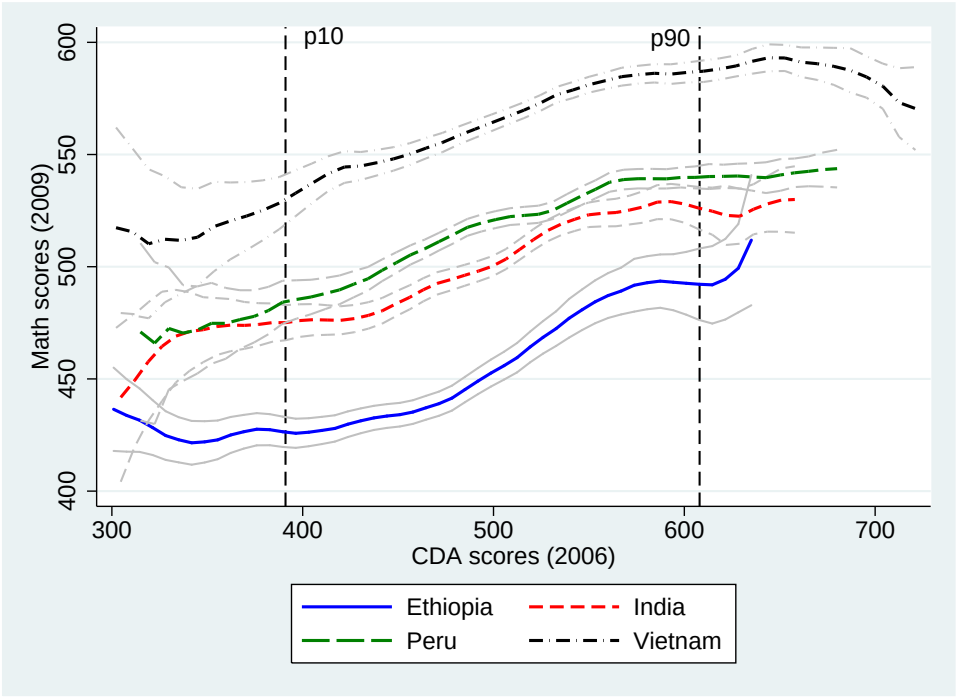
Figure 2: Learning distributions at different ages



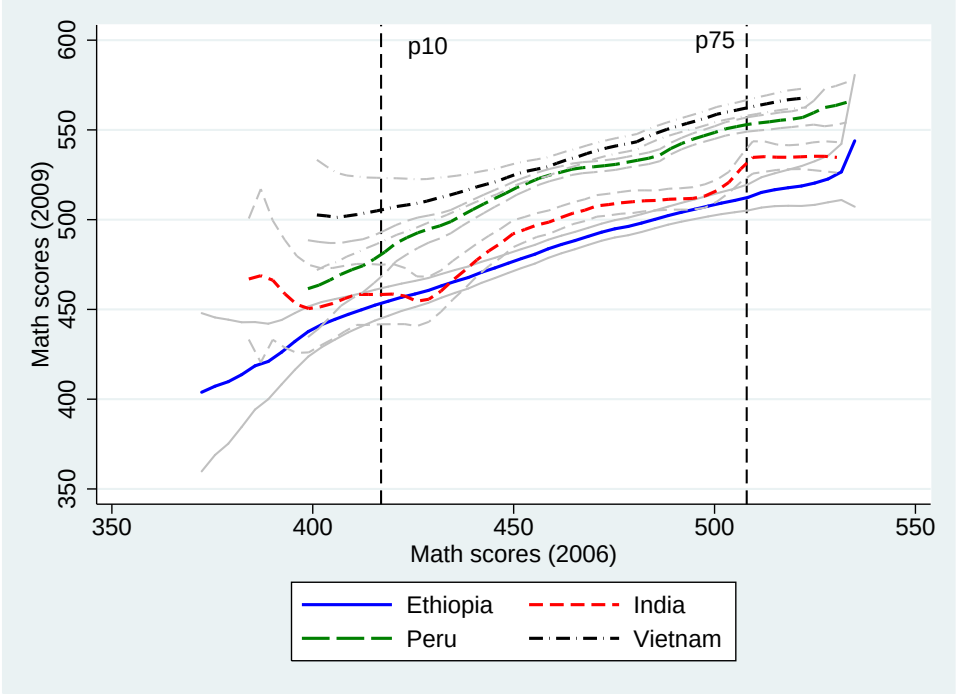
CDFs show distribution of test scores estimated with Item Response models pooling all country samples in each age group. Scores are internally normalized to have a mean of 500 and standard deviation of 100 in each age sample.

Figure 3: Progress in learning across countries

(a) 5-8 years



(b) 12-15 years



Note: Lines are local polynomial smoothed lines shown with 95% confidence intervals 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

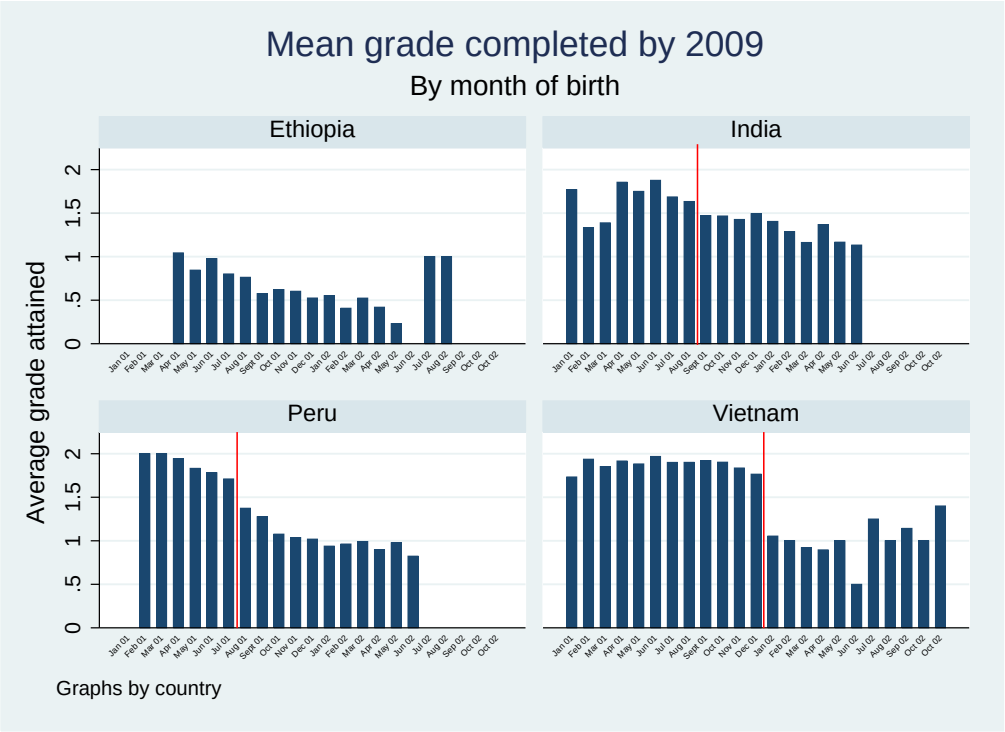


Table 1: Enrolment and grade progression of children in Young Lives

Cohort	Variable	Statistics	Ethiopia	India	Peru	Vietnam
Younger cohort (YC)	Age of starting school	Mean	6.9	5.78	6.04	6.05
		SD	0.85	0.82	0.42	0.27
Older cohort (OC)	Age of starting school	Mean	7.19	5.04	5.88	6.07
		SD	1.52	0.71	0.57	0.48
YC 2006 (5-years)	Enrolment	Mean	0.04	0.45	0.01	0.01
		SD	0.19	0.5	0.1	0.08
YC 2009 (8-years)	Enrolment	Mean	0.77	0.99	0.98	0.98
		SD	0.42	0.1	0.13	0.13
OC 2006 (12-years)	Enrolment	Mean	0.95	0.89	0.99	0.97
		SD	0.22	0.32	0.1	0.18
OC 2009 (15-years)	Enrolment	Mean	0.89	0.77	0.92	0.77
		SD	0.31	0.42	0.27	0.42
YC 2009 (8-years)	Highest grade completed	Mean	0.64	1.63	1.31	1.71
		SD	0.77	1	0.58	0.57
OC 2006 (12-years)	Highest grade completed	Mean	3.17	5.61	4.91	5.57
		SD	1.68	1.25	1.11	0.94
OC 2009 (15-years)	Highest grade completed	Mean	5.55	8.15	7.72	8.29
		SD	2.05	1.73	1.31	1.25

Younger cohort ('YC') were born in 2001/2002. Older cohort ('OC') were born in 1994/95
Age of starting school is summarized in both cohorts over those individuals who have enrolled in school at some point before the survey round in 2009.

Table 2: Proportion answering TIMSS questions correctly

Item details	Q1	Q2	Q3	Q4	Q5	Q6
TIMSS Item Number	M011007	M011024	M011011	M011004	M031310	M031162
Cognitive Domain	Using Concepts	Knowing Facts and Procedures	Solving Routine Problems	Using Concepts	Solving Routine Problems	Using concepts
TIMSS						
Canada - Quebec	0.925	0.926	0.847	0.894	0.690	0.636
England	0.929	0.959	0.875	0.858	0.787	0.818
Hong Kong	0.977	0.983	0.916	0.848	0.950	0.746
Italy	0.924	0.968	0.831	0.850	0.731	0.785
Japan	0.968	0.988	0.928	0.895	0.893	0.900
Morocco	0.699	0.775	0.457	0.475	0.441	0.491
Netherlands	0.976	0.949	0.949	0.896	0.893	0.864
Philippines	0.633	0.859	0.573	0.600	0.282	0.416
Singapore	0.974	0.970	0.936	0.895	0.941	0.878
Tunisia	0.813	0.739	0.609	0.495	0.622	0.376
USA	0.929	0.936	0.884	0.887	0.674	0.666
Young Lives						
Ethiopia	0.606	0.716	0.512	0.504	0.404	0.574
India (A.P.)	0.737	0.822	0.603	0.682	0.385	0.711
Peru	0.701	0.912	0.679	0.769	0.508	0.650
Vietnam	0.843	0.938	0.759	0.687	0.751	0.845

Questions focused on the number content domain and were taken from the released items for the 2003 TIMSS assessment.
The TIMSS sample is from Grade 4, aged 10 years on average, at the time of testing in 2003. Cells contain unweighted proportions.
The Young Lives sample is for the older cohort, aged about 12 years at time of testing in 2006.

Table 3: Comparing 12-year old children in Young Lives to TIMSS 4th grade sample

Country	Median grade	Median score	% below low benchmark (400)	% below intermediate benchmark (475)
Young Lives sample (12 years old, 2006)				
Ethiopia	4	394	52	82
India	7	460	28	54
Peru	6	454	25	59
Vietnam	7	525	13	38
TIMSS 4th grade sample (selected countries, 2003)				
Australia	4	504	12	36
Canada - Quebec	4	509	6	30
Chinese Taipei	4	567	1	8
England	4	536	7	25
Hong Kong	4	578	1	6
Italy	4	507	11	34
Japan	4	568	2	11
Netherlands	4	542	1	11
Russian Federation	4	533	5	24
Singapore	4	601	3	9
United States	4	522	7	28

Test scores are IRT scores linked to TIMSS 2003 assessment using item parameters of publicly released items for anchoring and normalized as in TIMSS.
TIMSS normalizes scores to have a mean of 500 and standard deviation of 100 in the international pooled sample. TIMSS sample performance taken from Mullis et. al. (2004).
Test scores are importantly affected in this sample by ceiling and floor effects: about 33% of children in Vietnam, 18% in India, 8% in Peru and 4% in Ethiopia answered all ten questions correctly and for these children the score is defined by the ceiling of 1000. About 8% of children in Ethiopia, 6% in India, 3% in Vietnam and 1% in Peru answered no questions correctly. Scores are bounded above and below at [0,1000]. This affects the mean of the distribution but allows for unproblematic comparison of the median and proportion attaining benchmarks.

Table 4: Linked test scores at 5, 8 and 15 years

Age group	Statistics	Countries				Total
		Ethiopia	India	Peru	Vietnam	
5 years	Mean	454	498.3	520.4	524.7	499.8
	p25	402.4	442.1	462.9	472.6	442.9
	p50	456.1	491	511	522.7	495.3
	p75	503.8	540.7	569.1	575.6	550.8
	SD	102.1	94.8	97.6	89.1	99.9
	N	1846	1904	1893	1935	7578
8 years	Mean	419.1	495.9	518.2	563.6	500
	p25	377.7	458.2	491.4	535.9	456.5
	p50	426	502	530	570	518.7
	p75	488.8	542.3	555.8	600.5	559.4
	SD	100.7	84.6	68.3	85.3	100
	N	1885	1930	1943	1964	7722
15 years	Mean	442.7	482.1	527.7	556	500
	p25	433.1	467	512.9	528	478.3
	p50	480.1	508.4	535.9	561.6	522.6
	p75	512	543.1	554.6	590.7	555.7
	SD	106.5	99.6	54.9	79.7	100
	N	974	977	678	972	3601

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 5: Descriptive statistics of control variables

	Ethiopia			India			Peru			Vietnam		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Panel A: Younger cohort												
Child and background characteristics (X_{ic})												
Male	0.53	0.5	1881	0.53	0.5	1903	0.5	0.5	1892	0.51	0.5	1916
First born	0.23	0.42	1881	0.39	0.49	1903	0.37	0.48	1892	0.46	0.5	1916
Caregiver's Education	2.95	3.73	1874	3.7	4.44	1900	7.75	4.64	1892	6.88	3.83	1908
Age in months	97.48	4.05	1879	96.03	3.92	1903	95.35	3.63	1890	97.09	3.75	1915
Height-for-age z-score	-1.21	1.05	1877	-1.44	1.03	1898	-1.14	1.03	1890	-1.07	1.05	1900
Wealth index (2006)	0.28	0.18	1881	0.46	0.2	1902	0.47	0.23	1892	0.51	0.2	1914
Time use (hours spent on a typical day; $TU_{ic,a}$)												
— Doing domestic tasks	1.66	1.37	1881	0.33	0.58	1903	0.87	0.7	1887	0.54	0.66	1899
— Tasks on family farm/business etc.	1.5	2.22	1880	0.01	0.1	1903	0.25	0.66	1886	0.09	0.48	1897
— Paid work outside household	0.01	0.28	1880	0.01	0.2	1903	0	0.08	1887	0	0.07	1897
— At school	4.91	2.54	1881	7.72	0.95	1903	6.02	0.9	1887	5.04	1.31	1898
— Studying outside school time	0.99	0.89	1881	1.86	1.09	1903	1.87	0.83	1886	2.82	1.49	1897
— General leisure etc.	4.44	2.39	1881	4.71	1.54	1903	4.13	1.65	1887	5.55	1.65	1898
— Caring for others	0.83	1.21	1881	0.21	0.5	1903	0.48	0.88	1886	0.24	0.66	1878
Panel B: Older cohort												
Child and background characteristics (X_{ic})												
Male	0.51	0.5	971	0.49	0.5	976	0.53	0.5	664	0.49	0.5	972
First born	0.2	0.4	971	0.31	0.46	976	0.31	0.46	664	0.37	0.48	972
Caregiver's Education	2.93	3.49	967	2.86	4.05	976	7.27	4.57	663	6.77	3.85	971
Age in months	180.34	3.58	971	179.76	4.24	975	179.1	4.1	661	181.12	3.83	972
Height-for-age z-score	-1.37	1.28	968	-1.64	1	970	-1.48	0.9	657	-1.43	0.91	967
Wealth index (2006)	0.3	0.17	971	0.47	0.2	976	0.52	0.23	664	0.52	0.19	970
Time use (hours spent on a typical day; $TU_{ic,a}$)												
— Doing domestic tasks	2.55	1.65	970	1.45	1.35	975	1.42	1.07	662	1.44	0.96	958
— Tasks on family farm/business etc.	1.34	2.09	970	0.49	1.72	975	0.68	1.49	662	1.05	2.13	958
— Paid work outside household	0.4	1.63	970	1.04	2.77	975	0.41	1.72	662	0.47	2	958
— At school	5.55	2.17	970	6.39	3.59	975	5.91	2.01	662	4.23	2.34	946
— Studying outside school time	1.84	1.23	970	2.01	1.54	975	2.09	1.12	662	3.06	2.13	941
— General leisure etc.	2.98	1.71	970	4.1	2.32	975	3.24	1.48	662	4.97	2.23	955
— Caring for others	0.67	0.93	970	0.28	0.75	975	0.73	1.18	662	0.16	0.64	951

Children in the older cohort were born in 1994-95 and children in the younger cohort 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 6: Do home factors and child-specific endowments explain divergence?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dep var: Mathematics score (2009)							
VARIABLES	8-years old (Younger cohort)				15-years old (Older cohort)			
Country dummies								
India	76.3***	64.5***	61.6***	16.3***	39.4***	13.8***	7.38*	6.62
	(3.01)	(2.92)	(2.97)	(3.58)	(4.67)	(4.11)	(4.29)	(4.23)
Peru	96.7***	79.1***	65.2***	48.2***	85.4***	62.6***	47.4***	57.6***
	(2.75)	(2.71)	(2.69)	(2.85)	(4.00)	(3.53)	(3.76)	(3.81)
Vietnam	146***	127***	108***	92.2***	113***	66.0***	55.4***	69.5***
	(3.04)	(3.06)	(2.97)	(3.45)	(4.26)	(3.95)	(4.00)	(4.26)
Background characteristics								
Male			3.79**	4.94***			8.77***	11.1***
			(1.68)	(1.60)			(2.48)	(2.47)
First-born child			6.56***	4.81***			4.38*	2.61
			(1.68)	(1.61)			(2.63)	(2.50)
Caregiver's education level			2.99***	2.18***			1.12***	0.33
			(0.24)	(0.23)			(0.35)	(0.33)
Age in months			2.94***	2.80***			-0.51	0.033
			(0.23)	(0.22)			(0.32)	(0.30)
Height-for-age z-score (2009)			10.3***	7.66***			5.49***	4.95***
			(1.05)	(0.97)			(1.35)	(1.29)
Wealth index (2006)			74.6***	47.8***			67.1***	44.7***
			(5.36)	(5.08)			(7.70)	(7.38)
Time use (hours spent on a typical day)								
— doing domestic tasks				0.76				2.70*
				(1.31)				(1.47)
— doing tasks on family farm etc.				-2.07*				0.79
				(1.22)				(1.30)
— doing paid work outside hh				0.32				0.29
				(8.52)				(1.31)
— at school				13.6***				8.59***
				(0.99)				(1.16)
— studying outside of school time				13.2***				8.07***
				(0.93)				(1.19)
— general leisure etc.				1.64**				3.11***
				(0.82)				(1.09)
— caring for others				0.58				-0.60
				(1.25)				(1.73)
Lagged test score (2006)		0.27***	0.12***	0.10***		0.48***	0.41***	0.33***
		(0.011)	(0.0099)	(0.0095)		(0.016)	(0.016)	(0.014)
Constant	421***	301***	57.2**	0.10	443***	228***	331***	194***
	(2.37)	(5.51)	(22.5)	(23.8)	(3.41)	(8.32)	(57.7)	(57.0)
Observations	7,573	7,573	7,522	7,465	3,595	3,583	3,554	3,513
R-squared	0.285	0.352	0.450	0.514	0.197	0.397	0.441	0.504
	F-tests of equality of coefficients (p-value)							
India = Peru	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
India=Vietnam	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peru=Vietnam	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.00

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 7: Country-specific production functions of achievement: 8 years old

	(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	1.33	5.04	8.95***	-0.55	2.07	4.52	9.32***	-0.25
	(6.55)	(3.19)	(2.47)	(2.65)	(5.36)	(3.52)	(2.76)	(2.66)
Eldest	7.17	3.59	9.37***	7.74*	3.38	2.75	7.89**	8.26*
	(4.54)	(3.05)	(3.02)	(3.77)	(3.66)	(3.35)	(3.30)	(4.06)
Caregiver's education level	3.37***	2.91***	2.69***	3.73***	2.28***	2.25***	2.52***	2.34***
	(0.79)	(0.82)	(0.46)	(0.88)	(0.56)	(0.58)	(0.43)	(0.81)
Age in months	2.73***	1.96***	2.67***	4.07***	2.23***	1.95***	2.67***	4.16***
	(0.52)	(0.60)	(0.35)	(0.54)	(0.50)	(0.56)	(0.33)	(0.56)
Height-for-age (2009)	14.4***	9.93***	6.77***	10.3***	7.55***	8.88***	6.27***	7.33***
	(2.66)	(2.07)	(2.09)	(3.12)	(2.33)	(1.96)	(1.89)	(2.17)
Wealth index (2006)	174***	44.7	25.3***	91.7***	106***	21.5	25.2***	64.9***
	(27.7)	(26.6)	(7.86)	(27.5)	(18.4)	(19.7)	(7.99)	(20.2)
Time use (hours on a typical day)								
— doing domestic tasks					0.47	3.64	7.37***	-4.15
					(3.66)	(4.37)	(2.06)	(4.46)
— doing tasks on family farm etc.					0.67	-16.5***	-0.44	-24.7***
					(3.55)	(5.58)	(1.95)	(4.86)
— doing paid work outside hh					-4.85	22.0***	-5.49	15.5
					(8.81)	(6.56)	(4.95)	(9.32)
— at school					12.6***	22.4***	9.21***	5.16
					(3.57)	(2.64)	(2.98)	(4.84)
— studying outside school time					19.8***	19.9***	7.61***	4.65
					(3.76)	(5.03)	(1.52)	(3.42)
— general leisure etc.					1.26	5.36*	2.33*	-2.71
					(3.20)	(2.86)	(1.19)	(2.98)
— caring for others					2.11	0.96	2.00	-8.75
					(4.71)	(5.04)	(1.17)	(6.38)
Lagged CDA score (2006)	0.071**	0.15***	0.13***	0.12***	0.044*	0.14***	0.13***	0.089**
	(0.026)	(0.029)	(0.020)	(0.040)	(0.023)	(0.028)	(0.019)	(0.031)
Constant	77.9	212***	162***	47.4	61.7	-6.91	76.7**	56.2
	(47.6)	(59.5)	(31.5)	(55.8)	(63.5)	(74.0)	(32.8)	(61.8)
Observations	1,835	1,892	1,888	1,907	1,834	1,892	1,881	1,858
R-squared	0.255	0.177	0.282	0.309	0.374	0.280	0.312	0.353

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: Country-specific production functions of achievement: 15 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
Male	24.0***	22.0***	-6.11	-7.49	27.8***	19.8***	-0.90	-3.05
	(7.87)	(5.71)	(4.10)	(4.37)	(7.42)	(6.53)	(3.71)	(4.19)
Eldest	0.46	8.29*	6.53*	3.03	-1.75	6.88	6.79*	-0.065
	(9.40)	(4.16)	(3.37)	(3.60)	(8.78)	(4.11)	(3.37)	(3.73)
Caregiver's education level	-1.36*	2.22***	1.09***	2.88***	-1.33	0.79	0.80***	2.17**
	(0.77)	(0.73)	(0.34)	(0.96)	(0.80)	(0.62)	(0.28)	(0.91)
Age in months	-0.34	-0.86	0.46	-0.77	-0.12	-0.0082	0.52	-0.017
	(0.95)	(0.63)	(0.40)	(0.85)	(0.93)	(0.60)	(0.35)	(0.85)
Height-for-age (2009)	8.32**	3.72	4.61*	4.97*	6.73**	3.39	4.80*	4.55
	(3.07)	(3.03)	(2.55)	(2.77)	(2.78)	(2.88)	(2.46)	(2.97)
Wealth index (2006)	120***	46.5**	34.1***	85.1***	88.0***	28.1	22.5***	66.0**
	(29.1)	(19.5)	(8.76)	(24.9)	(28.4)	(17.1)	(7.75)	(24.0)
Time use (hours on a typical day)								
— doing domestic tasks					6.14*	5.37	3.44*	2.79
					(3.05)	(3.42)	(1.81)	(3.77)
— doing tasks on family farm etc.					3.30	-3.82	0.46	-0.71
					(3.72)	(2.61)	(1.98)	(2.44)
— doing paid work outside hh					4.02	-4.70**	2.23	0.78
					(3.84)	(2.20)	(2.40)	(1.63)
— at school					10.7***	5.46**	7.35***	6.33**
					(2.89)	(2.42)	(2.28)	(3.00)
— studying outside school time					17.6***	7.99***	5.38***	1.32
					(2.83)	(2.09)	(1.53)	(1.90)
— general leisure etc.					7.11*	1.03	1.40	-1.29
					(3.64)	(1.66)	(1.37)	(1.97)
— caring for others					5.41	-9.77**	1.74	-1.59
					(3.84)	(3.69)	(1.06)	(4.08)
Lagged math score (2006)	0.64***	0.49***	0.22***	0.26***	0.55***	0.37***	0.19***	0.22***
	(0.048)	(0.059)	(0.040)	(0.034)	(0.041)	(0.042)	(0.032)	(0.032)
Constant	184	357***	321***	499***	50.2	223*	262***	369**
	(177)	(114)	(74.8)	(142)	(176)	(123)	(64.5)	(140)
Observations	964	970	656	964	963	969	656	925
R-squared	0.317	0.369	0.304	0.348	0.369	0.495	0.390	0.397

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 9: 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	420.79	485.28	495.47	523.15	420.75	390.94	486.66	488.38
		(9.87)	(10.64)	(5.49)	(13.48)	(10.85)	(16.72)	(9.62)	(19.19)
Inputs	India	450.36	497.32	503.74	539.9	487.38	497.32	516.86	563.24
(X_{ic}; TU_{ica})		(11.54)	(9.59)	(4.97)	(11.02)	(10.39)	(9.87)	(7.99)	(14.79)
$Y_{ic,a-1}$	Peru	470.66	514.64	517.73	559.32	479.48	468.87	517.74	557.66
		(11.35)	(10.7)	(4.65)	(10.53)	(10.93)	(10.96)	(5.65)	(11.68)
	Vietnam	478.69	518.05	522.35	567.03	492.1	476.78	520.84	568.22
		(11.08)	(9.76)	(4.51)	(9.16)	(12.06)	(13.14)	(7.09)	(11.43)
		Older cohort (15-years) Coefficients (β_c)							
		Without time use				**With time use**			
		Ethiopia	India	Peru	Vietnam	Ethiopia	India	Peru	Vietnam
	Ethiopia	443.17	448.98	507.7	502.26	443.15	453.52	512.33	524.55
		(10.54)	(10.14)	(7.12)	(9.21)	(12.13)	(11.38)	(8.5)	(12.03)
Inputs	India	495.01	482.61	524.44	529.91	496.15	482.86	531.03	549.28
(X_{ic};TU_{ica})		(13.12)	(9.84)	(6.54)	(8.67)	(14.96)	(10.38)	(9.05)	(12.55)
$Y_{ic,a-1}$	Peru	493.1	493.53	529.74	546.1	483.25	481.28	529.74	557.88
		(12.65)	(9.86)	(6.04)	(8.74)	(14.14)	(10.68)	(7.25)	(11.58)
	Vietnam	525.34	515.18	542.1	557.05	521.01	504.53	535.76	558.18)
		(12.65)	(10.25)	(6.56)	(8.54)	(14.1)	(11.36)	(9.14)	(10.56)

Cells contain linear predictions of test scores using combinations of country-specific production function parameters (β_c), as estimated in Tables 7 and 8 for 8-year and 15-year olds respectively with country-specific input levels (X_{ic} and TU_{ic}) as in Table 5. Each row shows predicted values of mean achievement when applying, to a given country sample, different country-specific coefficients indicated in column headings. Results are shown for specifications with and without time use categories. Standard errors of predictions in parentheses.

Table 10: 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	40.9***	27.4***	33.6***	60.9***	28.4***	25.4***	32.6***	55.2***
	(4.67)	(2.03)	(3.60)	(14.6)	(4.48)	(1.62)	(3.55)	(10.9)
Male	3.26	12.7***	8.73***	1.65	4.44	11.6***	8.92***	1.62
	(5.61)	(3.05)	(2.22)	(2.39)	(4.82)	(3.13)	(2.47)	(2.66)
Eldest	3.97	5.74**	8.45***	6.63**	1.63	4.59	7.09**	7.18**
	(4.08)	(2.60)	(2.89)	(2.98)	(3.72)	(3.01)	(3.09)	(3.16)
Caregiver's education level	3.76***	2.40***	2.23***	3.16***	2.74***	1.86***	2.10***	2.18***
	(0.66)	(0.70)	(0.49)	(0.80)	(0.52)	(0.49)	(0.48)	(0.72)
Age in months	1.26**	0.51	-0.067	0.18	1.30**	0.60	0.0079	0.69
	(0.53)	(0.45)	(0.30)	(1.10)	(0.56)	(0.41)	(0.30)	(0.87)
Height-for-age (2009)	9.31***	5.38**	5.22**	7.14***	5.30**	4.79**	4.82**	4.81***
	(2.64)	(2.21)	(1.92)	(1.78)	(2.33)	(1.85)	(1.73)	(1.56)
Wealth index (2006)	151***	53.6**	17.6*	78.3***	105***	31.0*	18.1*	59.0***
	(25.9)	(23.8)	(8.80)	(20.9)	(18.8)	(17.8)	(8.91)	(19.0)
Time use (hours on a typical day)								
— doing domestic tasks					2.23	3.06	6.72***	-4.16
					(3.37)	(4.21)	(1.95)	(3.89)
— doing tasks on family farm etc.					1.41	-13.6***	0.12	-21.4***
					(3.44)	(3.39)	(1.61)	(5.34)
— doing paid work outside hh					-3.94	22.3***	-4.05	-2.56
					(7.86)	(7.50)	(3.84)	(7.02)
— at school					12.1***	21.2***	8.94***	3.78
					(3.36)	(2.54)	(2.90)	(4.09)
— studying outside school time					13.8***	17.7***	6.77***	2.03
					(3.72)	(4.88)	(1.71)	(3.09)
— general leisure etc.					2.12	4.53*	2.40*	-2.64
					(3.24)	(2.52)	(1.31)	(2.59)
— caring for others					3.29	1.74	1.96*	-7.01
					(4.65)	(4.70)	(1.05)	(4.77)
Lagged CDA scores (2006)	0.067***	0.13***	0.100***	0.065*	0.045*	0.12***	0.100***	0.049
	(0.023)	(0.027)	(0.021)	(0.032)	(0.022)	(0.027)	(0.020)	(0.030)
Constant	196***	306***	401***	354***	129*	97.6*	313***	333***
	(49.2)	(45.5)	(29.5)	(74.1)	(72.0)	(53.8)	(38.8)	(65.5)
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	0.370	0.458

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 11: Comparing effectiveness of a grade of schooling: 15-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
Male	25.8***	22.0***	-1.57	-5.77	26.9***	18.8***	0.75	-2.59
	(6.96)	(4.57)	(3.59)	(4.34)	(6.02)	(6.08)	(3.69)	(4.07)
Eldest	-0.74	3.38	3.62	3.19	-1.62	4.24	4.24	0.67
	(7.91)	(3.68)	(2.70)	(3.60)	(7.50)	(3.95)	(2.92)	(3.61)
Caregiver's education level	-1.32*	2.01***	0.62*	2.43**	-1.30*	1.02*	0.51	1.88**
	(0.71)	(0.55)	(0.34)	(0.91)	(0.72)	(0.54)	(0.29)	(0.86)
Age in months	-0.49	-1.30**	-0.98*	-1.02	-0.36	-0.57	-0.74*	-0.35
	(0.90)	(0.53)	(0.47)	(0.82)	(0.89)	(0.56)	(0.39)	(0.83)
Height-for-age (2009)	1.11	-0.61	1.90	3.48	1.09	0.54	2.23	3.47
	(2.89)	(2.81)	(2.10)	(2.79)	(2.76)	(2.78)	(2.01)	(2.88)
Wealth index (2006)	76.8***	34.2*	6.09	68.2***	65.2**	25.2	3.85	56.5**
	(24.7)	(18.7)	(8.04)	(23.7)	(23.9)	(16.8)	(7.88)	(23.3)
Time use (hours on a typical day)								
— doing domestic tasks					4.30	2.88	2.52	1.39
					(3.21)	(2.80)	(1.61)	(3.72)
— doing tasks on family farm etc.					4.07	-2.43	0.61	-0.57
					(3.35)	(2.41)	(1.60)	(2.58)
— doing paid work outside hh					4.09	-3.29	2.49	-0.19
					(3.31)	(2.05)	(1.90)	(1.67)
— at school					8.16**	3.38	5.29***	4.84
					(3.06)	(2.31)	(1.66)	(2.94)
— studying outside school time					10.5***	6.08***	3.13**	0.60
					(2.87)	(2.01)	(1.16)	(2.08)
— general leisure etc.					5.52	-0.067	1.65	-1.46
					(3.63)	(1.70)	(1.19)	(2.10)
— caring for others					7.13*	-9.65**	1.16	-2.86
					(3.63)	(4.07)	(0.96)	(4.05)
Lagged math scores (2006)	0.34***	0.34***	0.14***	0.23***	0.33***	0.31***	0.14***	0.20***
	(0.046)	(0.051)	(0.019)	(0.029)	(0.046)	(0.041)	(0.016)	(0.030)
Highest grade completed	21.7***	19.9***	17.8***	12.2***	19.1***	13.1***	15.1***	10.3***
	(2.14)	(1.72)	(3.48)	(2.67)	(2.31)	(1.37)	(3.09)	(2.77)
Constant	227	349***	491***	474***	122	267**	424***	375**
	(167)	(96.2)	(73.7)	(133)	(163)	(114)	(55.8)	(135)
Observations	964	970	656	964	963	969	656	925
R-squared	0.427	0.482	0.454	0.408	0.443	0.533	0.486	0.435

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 12: 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	20.1***	20.9***	47.3***	46.3***
	(7.61)	(7.96)	(7.49)	(7.16)
Male	9.43***	9.96***	1.34	1.56
	(2.39)	(2.63)	(2.36)	(2.46)
Eldest	8.04***	6.44**	5.14	6.36**
	(2.74)	(3.01)	(3.16)	(3.05)
Caregiver's education level	2.31***	2.14***	3.05***	2.41***
	(0.40)	(0.37)	(0.61)	(0.55)
Age in months	0.94	0.87	0.41	0.64
	(0.66)	(0.71)	(0.57)	(0.53)
Height-for-age (2009)	6.15***	5.59***	6.00***	4.18***
	(2.20)	(2.00)	(1.96)	(1.44)
Wealth index (2006)	29.7***	29.0***	40.2**	28.6**
	(7.67)	(7.84)	(16.2)	(13.4)
Time use (hours on a typical day)				
— doing domestic tasks		4.99***		-3.37
		(1.71)		(4.40)
— doing tasks on family farm etc.		-0.13		-15.1***
		(2.20)		(4.86)
— doing paid work outside hh		0.42		-2.32
		(4.20)		(6.37)
— at school		8.45***		7.79
		(2.88)		(4.89)
— studying outside school time		6.72***		9.54***
		(1.54)		(2.73)
— general leisure etc.		0.95		0.18
		(1.16)		(1.77)
— caring for others		2.08**		-4.37
		(0.96)		(4.07)
Lagged math scores (2006)	0.13***	0.12***	0.11***	0.088***
	(0.020)	(0.020)	(0.031)	(0.027)
Constant	290***	227***	375***	316***
	(58.2)	(69.2)	(55.5)	(60.2)
Observations	1,888	1,881	1,907	1,858
R-squared	0.366	0.393	0.481	0.504
Kleibergen-Paap F-statistic	108	110	113	152

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 gradec completed is treated as endogenous in this table and instrumented for using in each country a discontinuity arising from en

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.

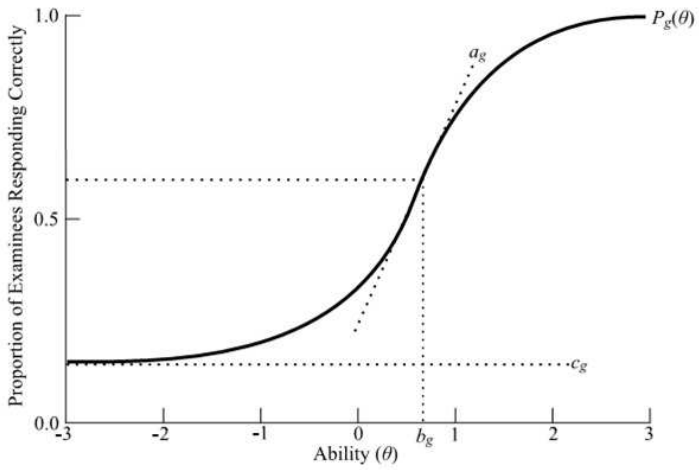
In this paper, following the procedure in TIMSS, I use the three-parameter logistic (3PL) model for all multiple-choice questions and the 2-PL model for items allowing open-ended responses. The 3-PL model is given by the following functional form:

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

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[37].

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

[37]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.

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 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[38].

[38]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.

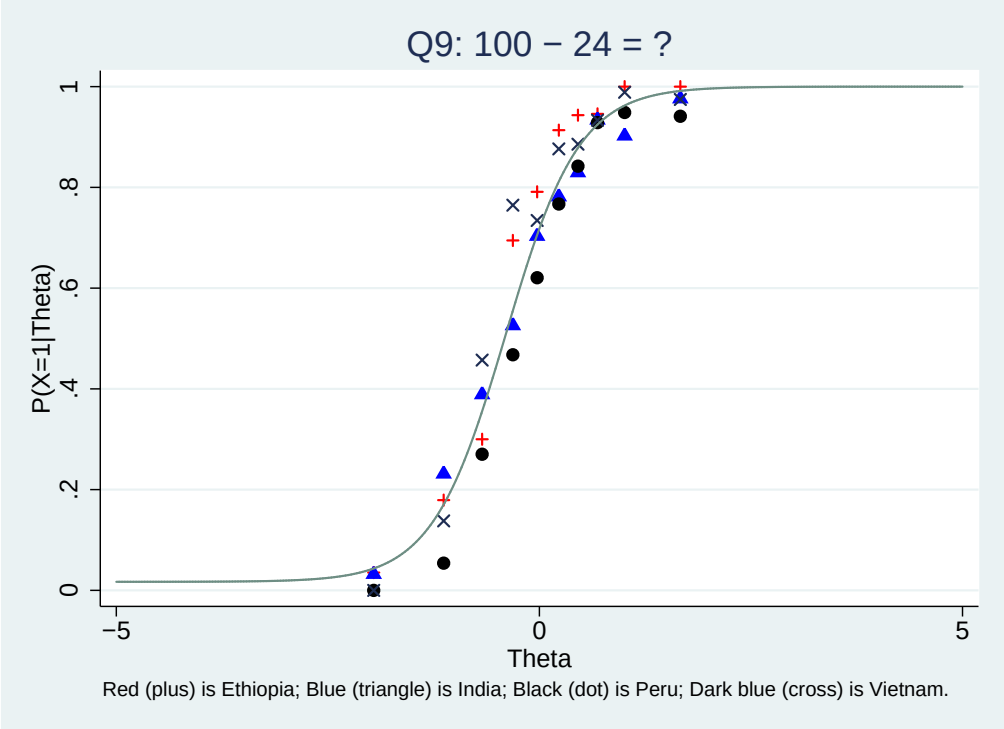
Table A.1: List of Items which were split in estimation due to DIF

Age sample	Item no.	Countries in which modified
5-years (CDA, 15 Items)	1	Ethiopia (D), India (S), Peru (D)
	3	Vietnam (S)
	6	India (S), Peru (D), Vietnam (D)
	7	Deleted 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)
12-years (Math, 10 items)	7	Ethiopia (S), Vietnam (S)
	8	India (S)
	9	India (S), Vietnam (S)
15-years (Math, 30 items)	12	India (S)
	14	India (S)
	21	India (S), Peru (S), Vietnam (S)
	22	Ethiopia (S)
	23	Ethiopia (S), India (S)
	26	Ethiopia (S), India (S)
	27	Ethiopia (S)
	28	Ethiopia (S)

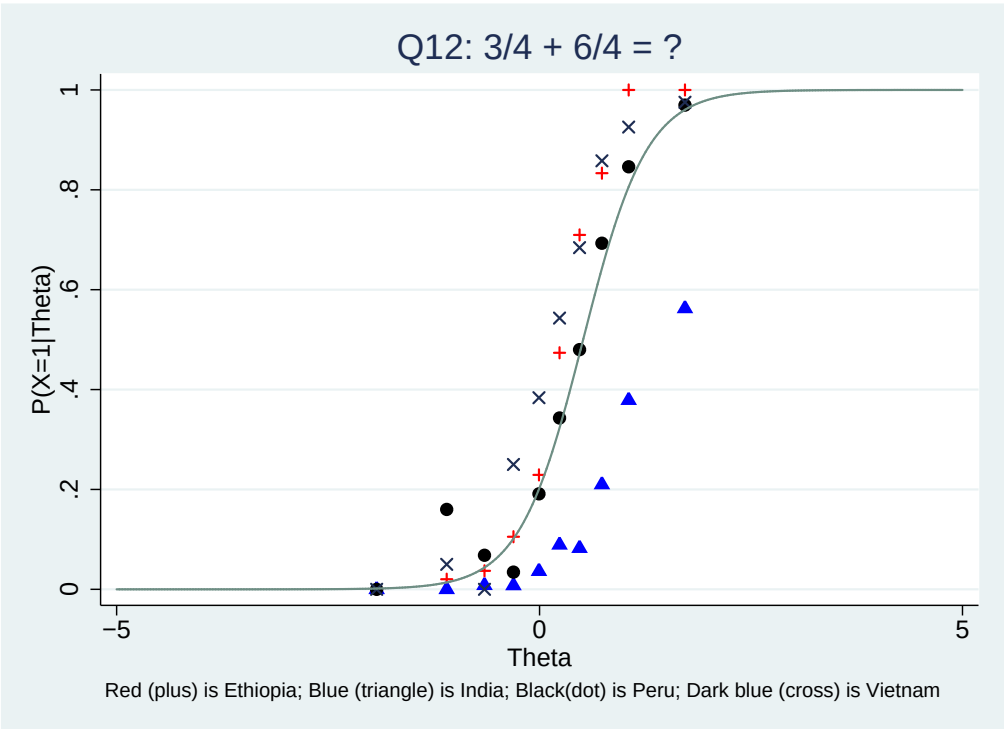
(D): Item deleted
(S): Item split in estimation

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

(a) No evidence of DIF



(b) Evidence of DIF



B Production function estimates with flexible lag specifications

Table A.2: 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
Male	2.89	12.5***	8.72***	1.43	4.15	11.3***	8.89***	1.45
	(5.52)	(3.08)	(2.28)	(2.36)	(4.84)	(3.14)	(2.51)	(2.58)
Eldest	4.10	5.19*	8.32***	6.55**	1.70	4.18	7.08**	7.23**
	(3.96)	(2.50)	(2.81)	(2.96)	(3.69)	(2.90)	(3.01)	(3.17)
Caregiver's education level	3.65***	2.36***	2.24***	3.06***	2.70***	1.82***	2.13***	2.17***
	(0.62)	(0.70)	(0.50)	(0.77)	(0.51)	(0.49)	(0.48)	(0.71)
Age in months	1.13**	0.41	-0.23	0.15	1.23**	0.52	-0.15	0.67
	(0.50)	(0.47)	(0.33)	(1.11)	(0.54)	(0.43)	(0.31)	(0.88)
Height-for-age (2009)	8.79***	5.42**	4.99**	7.02***	5.02**	4.80**	4.62**	4.65***
	(2.59)	(2.19)	(1.84)	(1.71)	(2.31)	(1.83)	(1.65)	(1.50)
Wealth index (2006)	148***	52.3**	14.4	73.8***	103***	30.0	15.3	56.4***
	(25.5)	(23.8)	(8.91)	(20.4)	(18.6)	(17.7)	(9.00)	(18.8)
Time use (hours on a typical day)								
— doing domestic tasks					2.15	3.04	7.07***	-3.84
					(3.34)	(4.16)	(1.96)	(4.06)
— doing tasks on family farm etc.					1.43	-13.6***	0.23	-21.3***
					(3.44)	(3.52)	(1.68)	(5.28)
— doing paid work outside hh					-3.57	22.1***	-5.39	-1.85
					(7.88)	(7.17)	(3.86)	(7.48)
— at school					11.9***	21.2***	8.93***	3.80
					(3.33)	(2.53)	(2.81)	(4.09)
— studying outside school time					13.8***	17.7***	6.38***	2.04
					(3.71)	(4.80)	(1.72)	(3.06)
— general leisure etc.					2.05	4.68*	2.36*	-2.47
					(3.23)	(2.49)	(1.27)	(2.57)
— caring for others					3.30	1.63	1.80	-7.02
					(4.69)	(4.71)	(1.06)	(4.80)
Highest grade completed	40.6***	27.1***	33.4***	60.0***	28.3***	25.1***	32.5***	54.8***
	(4.74)	(2.08)	(3.72)	(14.5)	(4.56)	(1.64)	(3.65)	(10.9)
Lagged test score	-0.89**	-0.43	0.32	-1.08**	-0.54	-0.43*	0.31	-1.08**
	(0.40)	(0.27)	(0.43)	(0.39)	(0.34)	(0.24)	(0.39)	(0.39)
Lagged score, squared	0.0020**	0.0013**	-0.000041	0.0024***	0.0012*	0.0012***	-0.000022	0.0022***
	(0.00081)	(0.00048)	(0.00071)	(0.00070)	(0.00070)	(0.00043)	(0.00064)	(0.00075)
Lagged score, cubed	-1.23e-06**	-9.16e-07***	-1.80e-07	-1.54e-06***	-8.02e-07*	-8.49e-07***	-1.87e-07	-1.36e-06***
	(5.19e-07)	(2.77e-07)	(3.87e-07)	(4.14e-07)	(4.48e-07)	(2.60e-07)	(3.53e-07)	(4.59e-07)
Constant	349***	385***	342***	532***	220***	177**	257***	520***
	(58.7)	(69.7)	(78.7)	(97.7)	(63.6)	(67.1)	(74.2)	(100)
Observations	1,835	1,892	1,888	1,907	1,834	1,892	1,881	1,858
R-squared	0.343	0.280	0.355	0.441	0.411	0.368	0.381	0.461

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 10 which is the analogous specification entering lagged achievement linearly

Table A.3: Allowing non-linearity in lagged achievement: 15-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
Male	24.6***	21.9***	-4.39	-6.01	26.4***	18.8***	-2.16	-3.08
	(7.06)	(3.24)	(3.35)	(4.24)	(6.35)	(4.67)	(3.52)	(4.00)
Eldest	-1.54	2.82	2.70	2.62	-2.47	3.60	3.55	0.61
	(7.59)	(3.72)	(2.93)	(3.49)	(7.20)	(4.01)	(3.08)	(3.60)
Caregiver's education level	-1.17*	2.00***	0.70**	2.17**	-1.17*	1.16**	0.59**	1.72**
	(0.65)	(0.49)	(0.28)	(0.87)	(0.66)	(0.54)	(0.26)	(0.82)
Age in months	-0.33	-0.35	-0.65*	-0.93	-0.19	0.18	-0.41	-0.29
	(0.90)	(0.46)	(0.36)	(0.79)	(0.89)	(0.46)	(0.29)	(0.79)
Height-for-age (2009)	1.60	-0.027	1.69	2.39	1.60	0.78	2.18	2.63
	(2.89)	(2.64)	(1.68)	(2.63)	(2.78)	(2.54)	(1.65)	(2.81)
Wealth index (2006)	67.9**	35.2**	-5.67	63.0**	55.4**	27.5*	-8.87	53.1**
	(23.9)	(15.4)	(7.92)	(23.5)	(22.8)	(15.3)	(7.77)	(23.1)
Time use (hours on a typical day)								
— doing domestic tasks					4.56	3.08	2.00	1.17
					(3.18)	(2.35)	(1.32)	(3.52)
— doing tasks on family farm etc.					3.99	-1.00	1.25	0.0012
					(3.40)	(1.98)	(1.51)	(2.32)
— doing paid work outside hh					4.11	-1.52	1.44	-0.19
					(3.37)	(1.65)	(1.81)	(1.63)
— at school					8.51**	3.60*	4.64***	4.58
					(3.17)	(1.84)	(1.33)	(2.81)
— studying outside school time					10.7***	5.73***	3.48***	1.02
					(2.67)	(1.95)	(1.03)	(2.03)
— general leisure etc.					6.00	0.14	1.54	-1.20
					(3.61)	(1.46)	(1.20)	(2.15)
— caring for others					7.99**	-8.84**	0.0050	-2.58
					(3.71)	(4.02)	(1.13)	(3.71)
Highest grade completed	18.8***	12.4***	11.4***	9.75***	16.0***	7.41***	8.83***	8.22***
	(2.23)	(1.98)	(2.65)	(2.16)	(2.44)	(1.36)	(2.50)	(2.25)
Lagged test score	-6.96***	-3.27*	0.90	-2.14	-7.11***	-2.65	0.60	-0.71
	(1.84)	(1.76)	(2.06)	(3.77)	(1.97)	(1.60)	(2.09)	(3.49)
Lagged score, squared	0.017***	0.010***	0.00099	0.0068	0.017***	0.0089**	0.0016	0.0035
	(0.0038)	(0.0036)	(0.0039)	(0.0075)	(0.0040)	(0.0033)	(0.0039)	(0.0069)
Lagged score, cubed	-0.000012***	-8.48e-06***	-1.90e-06	-5.37e-06	-0.000012***	-7.30e-06***	-2.28e-06	-3.06e-06
	(2.49e-06)	(2.40e-06)	(2.44e-06)	(4.88e-06)	(2.66e-06)	(2.21e-06)	(2.44e-06)	(4.45e-06)
Constant	1,215***	517*	118	655	1,120***	373	99.5	358
	(343)	(287)	(356)	(631)	(340)	(277)	(361)	(568)
Observations	964	970	656	964	963	969	656	925
R-squared	0.445	0.563	0.547	0.437	0.462	0.599	0.575	0.456

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 11 which is the analogous specification entering lagged achievement linearly

C Robustness of estimates at 15 to censoring of achievement at age 12

Table A.4: Production function estimates at 15-years using Bayesian EAP lagged scores

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
Male	24.4***	20.9***	-3.32	-5.43	26.1***	18.4***	-1.13	-2.68
	(7.02)	(4.27)	(3.69)	(4.30)	(6.24)	(5.77)	(3.74)	(3.93)
Eldest	-1.48	2.25	2.70	2.38	-2.44	3.25	3.37	0.17
	(7.68)	(3.49)	(2.68)	(3.38)	(7.32)	(3.70)	(2.95)	(3.46)
Caregiver's education level	-1.34*	1.70***	0.51	2.03**	-1.34*	0.86	0.40	1.62*
	(0.67)	(0.50)	(0.35)	(0.88)	(0.69)	(0.51)	(0.31)	(0.83)
Age in months	-0.39	-0.98*	-0.80*	-1.07	-0.26	-0.36	-0.56	-0.45
	(0.90)	(0.49)	(0.42)	(0.81)	(0.88)	(0.51)	(0.33)	(0.81)
Height-for-age (2009)	1.62	-0.50	1.38	2.39	1.61	0.49	1.72	2.61
	(2.87)	(2.55)	(2.09)	(2.59)	(2.76)	(2.48)	(1.99)	(2.76)
Wealth index (2006)	69.1***	32.6*	-3.92	62.9**	57.0**	24.9	-6.05	53.5**
	(23.7)	(17.7)	(8.11)	(23.5)	(22.8)	(16.5)	(8.10)	(23.1)
Time use (hours on a typical day)								
— doing domestic tasks					4.25	3.08	2.37	1.05
					(3.14)	(2.76)	(1.43)	(3.69)
— doing tasks on family farm etc.					3.71	-2.02	0.66	-0.35
					(3.32)	(2.20)	(1.48)	(2.44)
— doing paid work outside hh					3.83	-3.20	2.20	-0.17
					(3.36)	(1.90)	(1.86)	(1.66)
— at school					8.10**	3.15	5.30***	4.32
					(3.10)	(1.99)	(1.57)	(2.87)
— studying outside school time					10.5***	4.72**	2.66**	0.65
					(2.72)	(1.88)	(1.09)	(2.12)
— general leisure etc.					5.62	-0.23	1.52	-1.42
					(3.57)	(1.61)	(1.15)	(2.15)
— caring for others					7.31*	-9.60**	0.85	-2.64
					(3.72)	(4.11)	(1.05)	(3.80)
Lagged math scores (2006)	0.34***	0.43***	0.22***	0.33***	0.34***	0.39***	0.22***	0.29***
	(0.035)	(0.047)	(0.023)	(0.035)	(0.034)	(0.040)	(0.021)	(0.035)
Highest grade completed	19.6***	16.8***	15.3***	10.5***	16.9***	10.9***	12.6***	9.01***
	(2.06)	(1.68)	(3.32)	(2.26)	(2.31)	(1.24)	(2.95)	(2.45)
Constant	226	274***	444***	442***	120	211*	377***	357**
	(167)	(91.8)	(62.9)	(130)	(167)	(103)	(45.3)	(131)
Observations	964	970	656	964	963	969	656	925
R-squared	0.441	0.532	0.494	0.432	0.457	0.574	0.525	0.451

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.
The dependent variable is the maximum likelihood IRT estimate as in all previous tables.
Lagged achievement (at the age of 12) is the Bayesian EAP test score which is more robust to ceiling and floor effects.
Coefficients should be compared to Table 11 which is identical but for using MLE scores as the lagged achievement measure.

D Correcting for measurement error in lagged achievement

Table A.5: 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	40.9***	27.4***	33.6***	60.9***	28.4***	25.4***	32.6***	55.2***
	(4.67)	(2.03)	(3.60)	(14.6)	(4.48)	(1.62)	(3.55)	(10.9)
Male	1.18	14.1***	8.97***	1.65	3.89	13.4***	9.43***	1.63
	(4.84)	(2.67)	(2.33)	(2.54)	(3.91)	(2.83)	(2.46)	(2.51)
Eldest	-1.57	2.36	7.19***	3.57	-2.17	2.13	5.88**	4.89*
	(3.75)	(2.87)	(2.35)	(2.58)	(3.99)	(2.95)	(2.48)	(2.79)
Caregiver's education level	2.55***	1.73***	1.58***	2.57***	1.98***	1.47***	1.51***	2.17***
	(0.49)	(0.50)	(0.32)	(0.57)	(0.44)	(0.46)	(0.31)	(0.64)
Age in months	0.59	-0.073	-0.43	-0.79	0.55	0.010	-0.43	-0.31
	(0.39)	(0.32)	(0.32)	(1.10)	(0.44)	(0.31)	(0.32)	(0.86)
Height-for-age (2009)	6.72***	3.41*	3.05**	4.50***	3.28*	3.59**	2.76**	3.24**
	(2.06)	(1.99)	(1.45)	(1.45)	(1.93)	(1.76)	(1.36)	(1.35)
Wealth index (2006)	60.8***	68.0***	18.4***	40.3***	41.9***	48.0***	18.4***	31.7**
	(14.2)	(16.9)	(5.22)	(14.8)	(12.2)	(14.6)	(5.47)	(13.5)
Time use (hours on a typical day)								
— doing domestic tasks					2.60	5.17	4.50***	-1.49
					(2.61)	(3.74)	(1.43)	(3.61)
— doing tasks on family farm etc.					0.69	-18.3***	1.76	-12.0**
					(2.83)	(4.05)	(2.31)	(4.68)
— doing paid work outside hh					-6.12	17.1***	3.25	5.11
					(7.28)	(6.33)	(5.36)	(11.2)
— at school					10.8***	16.5***	6.62***	8.73**
					(3.44)	(2.82)	(2.35)	(4.42)
— studying outside school time					12.7***	11.5***	6.07***	8.85***
					(2.81)	(2.53)	(1.15)	(2.29)
— general leisure etc.					1.01	2.01	1.02	1.61
					(2.46)	(1.44)	(1.01)	(1.67)
— caring for others					1.90	0.13	2.10**	-1.92
					(3.63)	(4.18)	(0.89)	(3.16)
Lagged CDA scores (2006)	0.15**	0.32***	0.27***	0.20***	0.11*	0.27***	0.26***	0.15**
	(0.069)	(0.045)	(0.027)	(0.065)	(0.057)	(0.044)	(0.030)	(0.064)
Constant	299***	217***	339***	426***	228***	94.7**	281***	345***
	(41.0)	(33.5)	(28.4)	(83.0)	(54.2)	(42.1)	(36.8)	(65.2)
Observations	1,821	1,821	1,848	1,708	1,820	1,821	1,842	1,662
R-squared	0.409	0.343	0.323	0.486	0.466	0.413	0.346	0.507
Kleibergen-Paap F-statistic	72.3	81.2	201	56.2	73.6	82.7	193	53.0

*** 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 adpated Peabody Picture Vocabulary test in 2006 to correct for measurement error. Coefficients should be compared to Table 10.