Starting together, growing apart:

Gender gaps in learning from preschool to adulthood in four developing countries*

Abhijeet Singh
University of Oxford

Sofya Krutikova
Institute for Fiscal Studies

September 19, 2015

Abstract

This paper studies gender gaps in multiple learning domains – quantitative skills, vocabulary and reading – for children from the age of 5 to 19 years of age using unique panel data from Ethiopia, India, Peru and Vietnam on two cohorts of children with linked test scores and detailed household-based information. To the best of our knowledge, this is the most extensive comparable panel-based investigation of this question over a long age range, covering multiple countries and test domains, and with extensive background information to enable an investigation of the sources of observed gaps. In all countries, both for math and receptive vocabulary, we find that there are no gender gaps prior to school entry (5 years); these gaps emerge later, widening particularly between the ages of 12 and 15 years, favouring boys in Ethiopia and India and favouring girls in Vietnam; subsequently, these learning gaps appear to mostly persist until early adulthood. Our analysis, both cross-sectional and over time, pays special attention to issues arising from selective enrolment, from the ordinality of test outcomes, and from the issue of decay in test scores. Finally, we investigate the sources of divergence between 12-15 years using panel-based value-added models with a rich set of covariates including past achievement, child health, time use, parental education and wealth, and the quality of schooling. In our most extensive specifications, we can explain between half and two-thirds of the cross-sectional gender gap in test scores but a substantial unexplained portion remains.

*This is a preliminary draft, for comments only. Please do not cite or circulate without permission of the authors.

1 Introduction

Gender-based gaps in education have long been of concern to economists and policy makers for several reasons: inequalities in human capital may directly translate into later inequalities in labour force participation, the nature of employment and wages; if arising from factors unrelated to productivity, such inequalities could indicate a misallocation of resources; and, perhaps most importantly, equality of opportunity regardless of gender or other social markers remains a valuable policy objective in its own right.[1]

In developing countries, especially in policy discussions but also most previous academic work, the core focus on gender-based inequalities in education has related to enrolment and grade progression through school. In these areas, considerable progress has been made in the past 15 years.[2] Over the same period, however, there has been growing realization that years of schooling may hide substantial systematic differences in the actual levels of skill development in children. Even conditional on being enrolled, there may be systematic gender differencs in factors determining learning and thus valid (and perhaps increasingly important) questions remain about the presence, extent and sources of gender gaps in test scores within and across countries.[3]

In this paper, we focus on three central questions relating to this area using unique panel data from four developing countries: Ethiopia, Andhra Pradesh state in India, Peru and Vietnam. First, is there a gender gap in achievement and how, if at all, does it differ in magnitude and direction across countries and different domains of learning such as quantitative and language abilities? Second, how do these gaps evolve in each domain over the course of childhood – at what ages do they first emerge, and do they then substantially decline or increase with age? Third, what are the proximate sources of these gaps – can their emergence perhaps be explained by observed differences in household investments, child endowments, time use in different activities or the quality of school attended? These questions are fundamental to understanding whether and where there is cause for concern in this area, the severity of such concern, and to identify the ages and dimensions in which interventions might most be required.

[1] For an early statement of these grounds of concern, see Mill (1869, Chapter 4)

[2] For example, the latest UNESCO Global Monitoring Report documents that "*gender disparty in primary enrolment has been substantially reduced since 1999, but not eliminated*" (UNESCO 2015, p.155). It further states: "*Countries where gender gaps have been reversed underline the dynamic nature of achieving gender parity. Careful analysis of these trends is needed to inform future policy*" (p.166).

[3] For instance a recent paper highlighting the sharp reduction in gender inequality in access to schooling notes: *"Schooling attainment, as measured by grades of school completed, does not necessarily accurately reflect the learning outcomes of children, particularly in contexts of social promotion to the next grade level and large variations in school quality and in family background…These differences may have implications for gender differences in learning despite the same level of schooling attainment if girls are likely to attend different types of schools than boys, tend to take different classes than boys, are treated differently than boys in the same classes, or are treated differently outside of school than boys are."* (Grant & Behrman, 2010)

For most developing countries, even reliable descriptive evidence about gender gaps in learning remains scarce, especially in any comparable form. This largely reflects a scarcity of suitable data on student achievement. Internationally comparable household-based surveys, such as the Demographic and Health Surveys or the Living Standards Measurement Studies, collect details on the enrolment and current grade of individual children in the household but do not administer tests of learning. Comparative international assessments of learning, such as PISA or TIMSS, cover few developing countries and are further stymied, as are all school-based surveys or national exams, by the exclusion of children who are not enrolled in schooling or absent on the day of the assessment. This selection is particularly a concern in contexts where both enrolment and attendance may vary systematically by gender.[4]

Data used in this paper were collected by the Young Lives study which has followed two cohorts of children in the four countries from 2002 to 2014. The data are uniquely suitable for our purposes and present several particular strengths. Foremost, they cover a long age range from 5 to 19 years, i.e. from preschool age to early adulthood, with comparable tests of achievement across countries and ages in multiple learning domains. They thus enable a comparative investigation of gender-based differences in the trajectories of skill development across countries, domains and periods of childhood. Since the data are collected through home visits of a birth cohort, they do not suffer from selection due to enrolment in schools or attendance on the day of testing. The panel dimension of the data allows for the analysis of learning dynamics, in particular the extent to which gender gaps observed at any particular age may be accounted for by differences in achievement that were already evident at earlier ages. It further allows for the estimation of value-added models of achievement which allow for a more robust identification of the sources of gender-based divergence in achievement than is possible cross-sectionally. Rich background information on factors affecting achievement allows for an assessment, within these value-added models, of several distinct channels through which a gender gap might emerge. Finally, the four countries represent very different cultural contexts, with important differences in gender-related attitudes and social norms, and thus are likely to provide a broad spectrum of gender-based differences in test scores in developing countries.

Our analysis most closely resembles that of Fryer & Levitt (2010) and Bharadwaj et al. (2012), for the US and Chile respectively, where they use panel data to measure the emergence and magnitude of a gender gap in mathematics, focusing particularly on early gaps. Both Fryer and Levitt (2010) and Bharadwaj et. al. (2012) additionally present a comparative assessments of gender gaps in multiple countries using cross-sectional data on international assessments from

[4]PISA refers to the Programme for International Student Assessment administered by the Organization for Economic Development (OECD) and tests the achievment of 15-year old students through school-based tests. TIMSS refers to the Trends in Mathematics and Science Study which tests Grade 4 and Grade 8 students in mathematics and science and is organized by the International Association for the Evaluation of Educational Achievement.

PISA and TIMSS. The comparative component of their analysis is, however, limited by their data sources which are constrained by their coverage of few developing countries outside Latin America, their exclusion of the population of school-aged children who are not enrolled or absent on the day of assessment, and by their cross-sectional nature which precludes the analysis of learning dynamics and constrains the investigation of the sources of any observed gender gaps.

We believe there are four important ways in which our work goes beyond previous work including Fryer and Levitt (2010) and Bharadwaj et. al. (2012). *Foremost*, we are able to present a comparative panel-based analysis of gender-based trajectories in learning for four countries not covered previously. Panel-based investigations, and analysis of gender gaps in early school years, in these two papers concentrates only on the US and Chile and not comparatively; nor are the four countries in our sample part of the (cross-sectional) comparisons based on PISA and TIMSS data in these papers. *Secondly*, we do not restrict our attention merely to gaps in mathematics but focus on a wider range of assessments including vocabulary and reading comprehension. This is important given the possibility of gender-based heterogeneity in the domains where learning gaps open up, at what ages and in which contexts; *a priori*, it is not clear that we would wish to consider only gaps in mathematics to the exclusion of language abilities which are perhaps equally foundational in the education of children.[5] *Third*, we cover a much longer age range than any previous paper, presenting comparable assessments of the extent of gender gaps from the ages of 5 to 19 years i.e. from prior to school entry and to early adulthood. *Fourth*, we play close attention to issues around the measurement of test scores, in particular to concerns about the inherent ordinality of scores, which render comparisons of gaps across samples and changes across ages fragile (see in particular Lang 2010 and Bond & Lang 2013) and to issues of decay in test scores. To the best of our knowledge, this is the most extensive such investigation of gender differences in learning in developing countries in terms of the age range covered, domains of learning and the range of possible channels that can be investigated in value-added models of achievement to decompose these gaps.[6]

[5]Niederle and Vesterlund (2010) raise a similar query on why the (pro-boy) gender gap in mathematics in the US gets considerable attention but the (pro-girl) gap in verbal scores does not. They hypothesise that it may be because math scores predict wages and labour market outcomes but verbal scores do not. The basis for making such a judgment about which domains of learning to prioritize does not exist in our study countries — even correlational evidence on the returns to particular academic abilities in the labour market remains scarce in these contexts. Moreover, even if similar correlations were documented in developing countries, it is not obvious we would want to focus on only a single domain, whether mathematics or language abilities; if verbal abilities are viewed as intrinsically valuable, and expanding the range of capabilities of individuals, then gender gaps might still be worthy of investigation even if the correlation with wages or employment is low or insignificant.

[6]One of us has previously analyzed descriptive patterns in a subset of the data presented here, specifically for children at the ages of 8, 12 and 15 years, in Dercon and Singh (2013) as part of a wider analysis of gender gaps in multiple domains which additionally includes analysis of gender gaps in nutrition, subjective well-being, various socio-emotional measures and parnetal and child aspirations. The analysis in this paper, while building on this previous work, is substantially distinct in covering a broader age range using additional rounds of data, paying much more careful attention to aspects of measurement and ordinality and focusing much more deeply on the dynamics of learning acquisition. Dercon and Singh (2013) presented cross-sectional gaps at these ages with

Four patterns stand out from our analysis. First, in no country do we find any evidence of significant gender gaps in learning in either quantitative ability or receptive vocabulary at preschool age (5 years); gender gaps do, however, develop at later ages in most countries and are particularly evident after the age of 12 years, i.e. in a period coinciding with adolescence and post-primary schooling. Second, there is important heterogeneity in the direction of gender gaps - whereas significant gender gaps, where they do exist, typically favour boys in Ethiopia, India and Peru, they typically favour girls in Vietnam. Third, in contrast to most developed countries, we typically find lower evidence of heterogeneity across domains of learning in our study countries: where significant, gender gaps in mathematics and receptive vocabulary are consistently in the same direction. Finally, although we document significant differences in household investment, enrolment and other factors determining learning between boys and girls in the age groups at which gender gaps in learning emerge, these along with prior test scores in the relevant domain and the quality of schools enrolled in, only partially account for the emergence of gender gaps; the extent to which we can explain the emergence of the gaps with these observable characteristics and test scores differs across countries with the greatest explanatory power being in Andhra Pradesh (India) sample. Typically, we can explain between half and two-thirds of the cross-sectional gender gap at 15. The inability of rich background data to be able to account for a substantial portion of the gender gap is similar to evidence from developed countries.

These patterns have important implications. Foremost, they signal that whereas educational outcomes for boys and girls show few signs of systematic bias at primary school ages, gender gaps are frequently prominent after the age of 12 years and, having emerged, often persist until adulthood: this indicates that interventions seeking to reduce gender gaps in eventual adult cognitive outcomes may best be targeted around the age of 12-15 years, a period marking the important transitions into post-primary education and into adolescence. Secondly, noting the contrast between Vietnam (where gender gaps typically favour girls) and the other countries particularly India and Ethiopia, it seems that the presumption of gender gaps necessarily favouring boys (as often implicit in policy discussions) is not always borne out by the data. in these cases, education policy needs to be responsive to the direction of gender gaps.[7]Thirdly, we note that the magnitude of gender gaps is often small and statistically indistinguishable from zero; these gaps are also considerably smaller than achievement gaps in other dimensions such as socio-economic status (see Krutikova and Singh 2015). This is in stark contrast to the widely-documented phenomenon of learning outcomes being very low in absolute terms in developing countries, with most children being below grade-level competence and often lacking

some cross-sectional decompositions but did not estimate value-added models to study learning dynamics; relied exclusively on raw scores thereby not making formal attempts to link test scores on a comparable metric; and did not include an analysis of reading ability.

[7]For an example of such policy focus, see for instance the United Nations Girls Education Initiative, the related £340 million Girls Education Challenge program funded by UK aid through the Department of International Development or the Nike Foundation's The Girl Effect program.

functioning skills despite years of formal schooling (see e.g. Glewwe & Kremer 2006; Pritchett 2013). In our judgment, this indicates that the broader problem of a 'learning crisis' in developing countries is considerably more salient and pressing than gender differences in learning.

The rest of this paper is structured as follows. Section 2 presents details on the data used in this paper including the tests administered and the procedures undertaken for lining them on a common metric and the analytical methods to handle the problem of ordinality of measures. Section 3 presents descriptively the extent of gender gaps in various domains and examines further the patterns of divergence over ages and by various levels of prior achievement. Section 4 investigates the sources of these gender gaps, relying on panel-based decompositions and value-added models to understand the extent to which we can explain the observed gaps with observed characteristics including education expenditures, child nutrition, child time use, the quality of schools, and various background characteristics; Section 5 concludes.

2 Data

Data used in this paper come from the Young Lives study, based at the University of Oxford, which has followed two cohorts of children in four countries – Ethiopia, Andhra Pradesh state in India, Peru and Vietnam – over four waves between 2002 to 2013/14.[8] The older cohort, of about 1000 children each in the four countries, was born in 1994/95 while the younger cohort (about 2000 children in each country) was born in 2001/02. The ages of the children at the time of the various survey rounds is given in Figure 1. In this paper, we use data from Rounds 2-4 of the survey i.e. observing the younger cohort children at the ages of 5, 8 and 12 years and the older cohort children at 12, 15 and 19 years.[9] The survey tracks children who migrated in later rounds from their initial community in the 2002 round and attrition rates in the sample are very low with over 90% of the sample still in the 2013/14 round.[10]

Household data

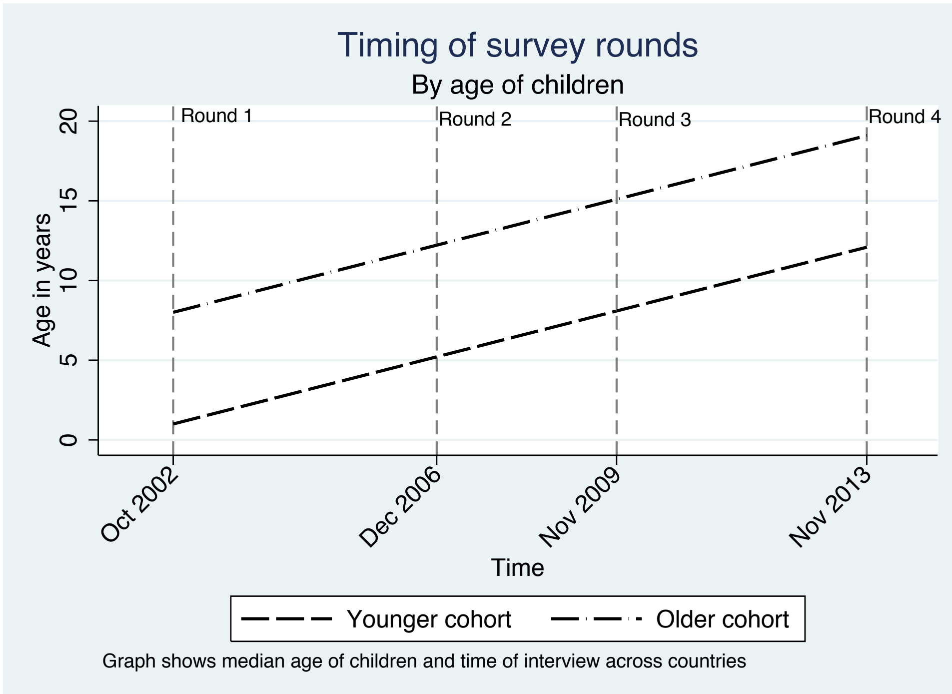
In each round of the survey, detailed questionnaires were administered regarding various household characteristics and child-specific information. This includes standard demographic

[8]Over the period of this study, the state of Andhra Pradesh (with a populaton of 84 million people in 2011) was bifurcated into Telangana and Andhra Pradesh states in 2014. Throughout this paper, when referring to Andhra Pradesh, we mean the undivided state as it existed until 2014. In terms of enrolment and learning outcomes, Andhra Pradesh is typically close to all-India averages (see e.g. Pratham 2015). In the paper we will often refer to results for 'India' or 'the Indian sample'; readers are requested to keep in mind that the sample is exclusively based in this one state.

[9]The first round of the survey administered only minimal assessments of learning to children in the older cohort, then aged 8. The younger cohort, then aged ~12 months, were not administered any cognitive assessments.

[10]Specifically, attrition from all causes excluding deaths between 2002 and 2013 rounds is under 5% for the younger cohort in all countries except Peru, where it is 6.3%. Attrition in the older cohort from all causes excluding deaths ranges from 4.3% in Andhra Pradesh to 11.3% in Vietnam. A detailed breakdown of attrition from the sample is available on the Young Lives website at www.younglives.org.uk.

Figure 1: Age of Young Lives sample individuals in successive survey rounds



and socio-economic information such as household structure, parental education, access to services and wealth but also, importantly for the purpose of this paper, extensive information on the individual child including time use, expenditures on the index child's education and their nutritional status (measured using WHO anthropometric scores). This information, combined with the detailed test score data in multiple rounds allows for a much deeper analysis of the sources of any gender gaps than has heretofore been possible in the literature for developing countries.

Tests administered in Young Lives

The Young Lives study has administered a wide variety of tests in various rounds which are summarized in Box 1. The tests differ in their mode of administration and the domain of learning they seek to assess. Quantitative skills are assessed at the age of 5 years using the orally administered quantitative sub-scale of the Cognitive Development Assessment (CDA) for pre-school aged children. At later ages, i.e. at 8, 12, 15 and 19 years of age, quantitative skills are assessed using paper-based mathematics tests. Given wide variations in the grade and skill levels of individuals both within and across countries, the tests are not designed to be grade-appropriate and incorporate in each round questions at widely differing levels of difficulty. At the lower end of difficulty, this includes questions assessing basic number recognition and computation whereas at the older end, it includes applying mathematical reasoning and procedures to applied problems.

Language skills are measured using a battery of different tests over the course of the study. Receptive vocabulary is tested using adapted versions of the Peabody Picture Vocabulary Test in the four countries between the age of 5 and 15 years; this is the language test most consistently administered across rounds and ages in the study and the primary focus of the analysis here. At the age of 15, in the third round of the survey (2009), language skills were additionally tested using a cloze test where students need to complete a sentence by providing a missing word ('fill-in-the-blanks'). Reading comprehension is measured in the fourth round of the survey (2013/14) using a language-specific reading comprehension test delivered to 12 and 19 year olds. [11]

Box 1. Cognitive Tests in Young Lives

COHORT	ROUND 2 (2006)	ROUND 3 (2009)	ROUND 4 (2013)
Younger Cohort	**5 years old**	**8 years old**	**12 years old**
	Receptive vocabulary CDA Quantitative	Receptive vocabulary Mathematics	Receptive vocabulary Mathematics Reading
Older Cohort	**12 years old**	**15 years old**	**19 years old**
	Receptive vocabulary Mathematics	Receptive vocabulary Mathematics Cloze test	Reading Mathematics

CDA refers to the Cognitive Development Assessment quantitative sub-scale.

Item Response Theory test scores

Test scores used in this paper are constructed using Item Response Theory (IRT) models which are commonly used in international assessments such as PISA and TIMSS. There are three key advantages for using IRT scores in this paper. Most importantly, using common items administered across age/round/country samples as 'anchor items' for the linking of tests scores, it is possible to put scores for different samples on the same metric; most obviously, this enables an assessment of whether the absolute magnitude of gender gaps increases and decreases with

[11] Please see Appendix A for further details on the tests available in Young Lives and the methods employed for construction of IRT test scores including various checks for differential item functioning across countries.

age or time or whether it is larger or smaller in one context compared to another. Secondly, Item Response models allow for a better range of diagnostics to assess the comparability of assessments across contexts or time. This is important because students in particular contexts, particular ages, or at some particular point in time, may be more familiar with certain types of questions or specific modes of testing. This is not a first-order issue for comparing gender gaps within one sample but is central to being able to compare across ages, time or countries. Finally, by allowing test questions to differ in their characteristics such as difficulty, IRT scores provide a less arbitrary measure of achievement than commonly reported percentage correct scores. This is important because it is possible that gender gaps in achievement are particularly concentrated on questions of particular difficulty levels: aggregation of test scores which provide an equal weight to all questions may then provide a misleading estimate of gender gaps.[12]

IRT models only identify ability up to a linear transformation and therefore require normalization. In this paper, we normalize scores to have a mean of zero and a standard deviation of one in the base age group in which it is administered. Specifically, mathematics scores are normalized with reference to the distribution of test scores of 8-year olds pooled across countries; the receptive vocabulary scores are normalized with reference to the 5-year old age group within language; and the reading scores, available only for 12 and 19 year olds, are normalized with reference to the 12-year old age group within-language.

Ordinality of test scores

A key issue, common to all test scores, is that of ordinality; any monotonic transformation of test scores is conceptually an equally valid test score. This is not a problem solved either by the use of Item Response Theory models or by applying common procedures for standardization and, as Bond and Lang (2013) demonstrate with reference to the Fryer & Levitt (2004) analysis of Black-White test score gaps, can be a serious concern in comparing inter-group differences in test scores.

Ordinality of the outcome measure greatly complicates the study of inter-group differences, rendering results on the magnitude (and potentially even direction) of gaps suspect.[13] Bounding approaches, seeking to establish the direction and range of magnitudes observed under various

[12] As an example of how this may matter in practice, Singh (2015) documents that the causal private school effect on English scores in rural India looks considerably larger when using IRT scores, in the metric of standard deviations of the score, than a standardized raw score. Dividing questions by the task required, he further demonstrates that the private school effect is larger on 'harder' questions than on 'easier' ones, thus making issues of weighting particularly salient.

[13] Specifically, if the CDFs of two groups cross, then even the direction of the gap can be reversed by arbitrary rank-preserving scaling decisions. Where one group first order stochastically dominates the other, the direction of the gap is robust to ordinality.

flexible rank-preserving transformations of the score, are likely to provide bounds too wide to be informative.[14]

Our approach to this issue is pragmatic. First, we restrict our (cross-sectional) comparisons within sample not just to documenting differences in the mean of test scores but across the entire distribution. Specifically, we also test for equality of distributions and first order stochastic dominance which serves to assure that the existence of any documented gender gaps in mean test scores is invariant to rank-preserving transformations of the test metric. Second, in a panel setting when we are looking at the divergence of test scores across ages, we will in all cases investigate non-parametrically the differences in test score *trajectories* between boys and girls. Specifically, for any two successive age points, we will plot separate local polynomial regressions predicting scores in the later period based on initial scores. To the extent that these trajectories do not cross, i.e. one trajectory is always above the other, this would serve to assure that our conclusions about the age periods in which divergence occurs, rather than merely the existence of a cross-sectional gap, are invariant to rank-preserving transformations of the test scores. Third, we concentrate always on *levels* of test scores, which may be more easily treated as approximately on an interval scale, rather than changes in test scores where such an interpretation is less justified.[15]

3 Emergence and evolution of gender gaps in learning

In this section, we first present descriptive patterns about the existence, direction and magnitude of gender gaps in enrolment. Subsequently, we move to investigating cross-sectionally gender gaps in mathematics, receptive vocabulary and reading test scores in the two cohorts at various ages. Finally, we use the panel dimension of the data to document gender-based divergence at different levels of ability over time.

3.1 Gender gaps in enrolment and grade progression

As noted previously, enrolment and grade progression have been the main focus of policy audiences and much academic work around gender equality in education. Gender gaps in these

[14] In their application to the Black-White test score gap in the ECLS-K and the CNLSY datasets, Bond and Lang (2013) document that bounds they derive are consistent with a change in the Black-White test score gaps between kindergarten and third grade lying somewhere between 0 and 0.6 standard deviations.

[15] Bond and Lang (2013) provide the following quote from Thorndike (1966) which suggests lesser fragility of results to ordinality when considered in levels than in changes: "*...it is assumed that the numerals in which the variables are expressed represent equal increments in some attribute. It is also recognized that this assumption is not usually well-supported. But for 'rough-and-ready' studies of relationship, the violation of the assumption does not hurt much. However, when starting to deal with something as fragile as a change score, the violation of this basic assumption becomes a good deal more critical.*"

Table 1: Enrolment and grade progression by gender and age

Panel A: Proportion enrolled

Age	Year	Ethiopia			India			Peru			Vietnam		
		Female	Male	Diff	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff
5	2006	0.04	0.03	0.00	0.45	0.44	0.01	0.01	0.01	0.00	0.01	0	0.01
8	2009	0.78	0.75	0.03	0.99	0.99	-0.00	0.99	0.99	-0.00	1.00	1.00	0.00
12	2014	0.96	0.93	0.03**	0.97	0.97	-0.00	0.99	0.99	-0.00	0.98	0.97	0.01
12	2006	0.96	0.94	0.02	0.88	0.90	-0.03	0.99	0.99	0.00	0.97	0.97	0.00
15	2009	0.91	0.88	0.04	0.74	0.81	-0.07*	0.95	0.91	0.04	0.81	0.73	0.08**
19	2014	0.62	0.54	0.08*	0.40	0.53	-0.13***	0.49	0.51	-0.02	0.51	0.42	0.09*

Panel B: Highest grade completed

Age	Year	Ethiopia			India			Peru			Vietnam		
		Female	Male	Diff	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff
8	2009	0.66	0.62	0.04	1.82	1.57	0.25***	1.31	1.32	-0.01	1.73	1.70	0.03
12	2014	3.59	3.53	0.07	5.67	5.30	0.37***	5.02	5.02	-0.00	5.74	5.68	0.06
12	2006	3.25	3.15	0.10	5.60	5.60	-0.00	4.97	4.85	0.12	5.57	5.57	0.00
15	2009	5.70	5.30	0.39**	8.13	8.16	-0.03	7.88	7.79	0.08	8.32	8.20	0.12

two measures are thus of interesting in themsleves. Moreover, they are important as contextual information to later understand gaps in test scores and to understand the magnitude of selective exclusion that may plague school-based assessments of gender gaps in test scores at different ages. Table 1 presents some basic descriptives on the enrolment patterns and grade progression of boys and girls at various ages.

At the age of 5 years, in nearly all countries, children have not yet transitioned to formal schooling and are still at preschool age; the only exception is India, where about 45% of the sample have joined school already. At the age of 8 years, there are few signs of gender bias in enrolment at 8, with enrolment near universal for both boys and girls, except in Ethiopia where a significant proportion of children start schooling later. Rates of grade progression are similar for boys and girls in all countries except India, where girls have on average completed 0.3 grades more.[16]

This absence of gender gaps in enrolment is true also at 12 in all countries for children in both cohorts with the exception of the younger cohort in Ethiopia, where there is a slight pro-girl bias in enrolment. Importantly, at 12 years, enrolment is near-universal in all countries – the only exception to this is the older cohort in India (where about 10% of children had dropped

[16] Perversely, in the Indian context, this is a sign of greater gender bias. Specifically, as documented in Singh (2014), children who will eventually enrol in private schooling spend longer in kindergarten classes and start school later; given that boys are more likely to enrol in private schools than girls, this leads to a higher grade progressin for girls in India at younger ages.

out by 12) but even there, the younger cohort 7 years later has near-universal enrolment at the same age.

At the age of 15, a notable gender gap in enrolment is noticeable in India (favouring boys) and Vietnam (favouring girls), in both cases about 7-8 percentage points. Finally, by the age of 19, when children are in typically in higher secondary grades or college, gender-based differences in enrollment are prominent in all countries except Peru but differing in direction: enrolment at 19 is sharply biased favouring boys in India but in both Vietnam and Ethiopia favours girls.[17]

3.2 Cross-sectional gaps in test scores

The primary focus of this paper is on gender differences in test scores; although gender gaps in enrolment and grade progression seem small, at least till primary school ages, it is still possible that this seeming gender parity may hide significant variation in actual student achievement by gender.

As a first step, we present cross-sectional differences in scores at different ages in each of the tests for the two cohorts. In Table 2 we document the presence, extent and direction of gender gaps in test scores by presenting the mean scores for boys and girls in each country/age sample for each test alongside p-values for two-sided tests of the equality of means (t-tests) and the equality of distributions (Kolmogorov-Smirnov tests). Further, for any samples where we see a significant gender difference in means, we test for first-order stochastic dominance in the same direction as the difference-in-means; the p-values from these Wilcoxon rank-sum tests are also presented in Table 1.

In order to ease presentation, we graphically present these differences-in-means for each age group in Figure 2, displaying the coefficient on a male dummy from a regression of test scores in each country with 95% confidence intervals; standard errors are clustered at site level in each country.

The first important feature to note, looking at both quantitative skills and receptive vocabulary, is that there is no indication of a gender gap in mean test scores at the age of 5 years i.e. around school entry age in any country. Differences between the mean scores of boys and girls are invariably small, although a little larger in receptive vocabulary than in quantitative skills, and we are unable to reject the equality of means at the 5% level in any country. Equality of means could mask important differences in the spread of distributions. However, we find little evidence of any such differences in distributions: we cannot reject equality of distributions in

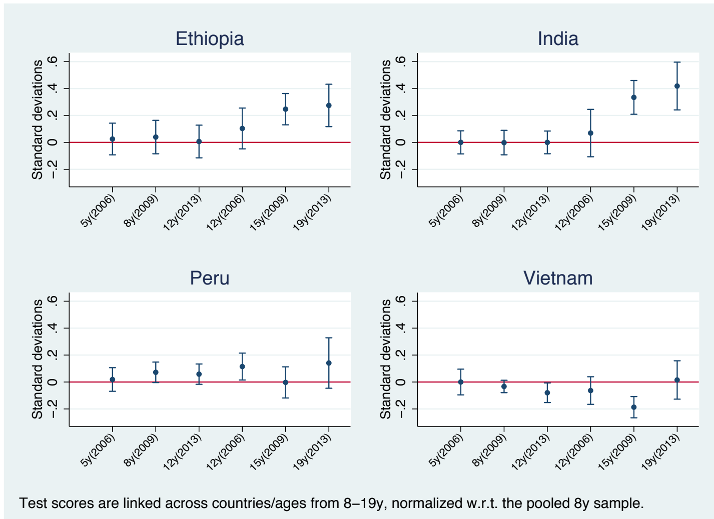
[17]The median enrolled child is enrolled in higher education (university or post-secondary technical) in India, Peru and Vietnam. In Ethiopia, due to a much-delayed age of starting school, the median enrolled child is in late secondary grades.

Table 2: Test scores of boys and girls from 5-19 years

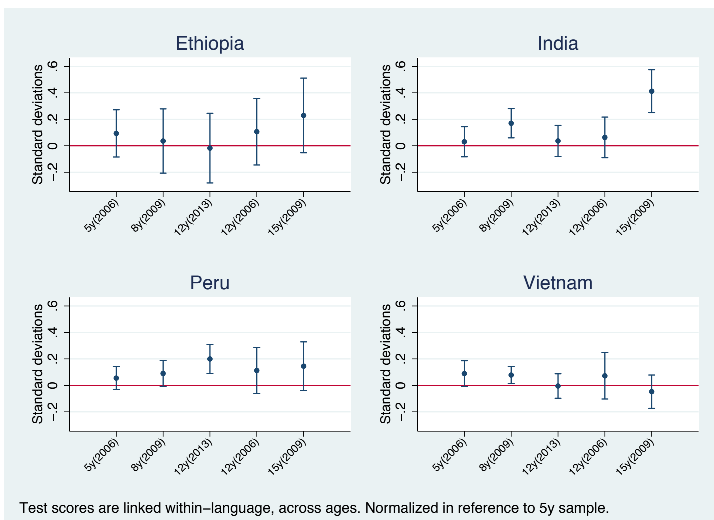
Domain	Age	Cohort	Ethiopia				India				Peru				Vietnam			
			M	F	t-test (p-value)	K-S test (p-value)	M	F	t-test (p-value)	K-S test (p-value)	M	F	t-test (p-value)	K-S test (p-value)	M	F	t-test	K-S test (p-value)
Quantitative skills	5	YC	-0.51	-0.54	0.58	0.69	-0.03	-0.04	0.99	0.70	0.22	0.20	0.65	0.78	0.32	0.32	1.00	0.80
Math	8	YC	-0.88	-0.93	0.34	0.27	-0.01	-0.01	0.98	0.84	0.20	0.13	0.04	0.05	0.68	0.71	0.30	0.11
	12	YC	-0.12	-0.13	0.88	0.25	0.35	0.35	1.00	0.55	0.81	0.76	0.08	0.10	1.43	1.51	0.03	0.03
	12	OC	0.37	0.26	0.07	0.12	0.89	0.82	0.24	0.08	0.91	0.79	0.05	0.04	1.29	1.35	0.21	0.31
	15	OC	0.13	-0.11	0.00	0.00	0.97	0.63	0.00	0.00	1.18	1.18	0.95	0.84	1.61	1.81	0.00	0.01
	19	OC	0.80	0.53	0.00	0.00	0.93	0.51	0.00	0.00	1.31	1.16	0.07	0.03	1.81	1.79	0.82	0.66
Receptive vocabulary	5	YC	0.04	-0.05	0.16	0.03	0.01	-0.02	0.52	0.16	0.03	-0.03	0.23	0.04	0.04	-0.05	0.05	0.18
	8	YC	1.29	1.25	0.70	0.36	0.84	0.67	0.00	0.00	1.72	1.63	0.05	0.45	1.73	1.65	0.04	0.00
	12	YC	2.21	2.21	0.84	0.42	1.94	1.91	0.41	0.53	3.26	3.06	0.00	0.00	3.01	3.01	0.91	0.16
	12	OC	2.29	2.18	0.22	0.13	2.40	2.34	0.35	0.04	2.49	2.37	0.09	0.46	3.61	3.54	0.38	0.64
	15	OC	2.83	2.60	0.03	0.04	2.77	2.35	0.00	0.00	3.85	3.71	0.06	0.02	3.48	3.54	0.51	0.24
Reading	12	YC	-0.05	0.05	0.12	0.13	-0.04	0.05	0.05	0.03	-0.01	0.01	0.69	0.48	-0.13	0.14	0.00	0.00
	19	OC	0.55	0.59	0.62	0.73	0.80	0.51	0.00	0.00	0.87	0.83	0.61	0.85	0.51	0.78	0.00	0.00
Cloze	15	OC	0.01	-0.01	0.89	0.31	0.12	-0.12	0.00	0.00	-0.02	0.02	0.60	0.41	-0.12	0.12	0.00	0.00

Figure 2: Mean gender differences in achievement from 5-19 years: Coefficient plots

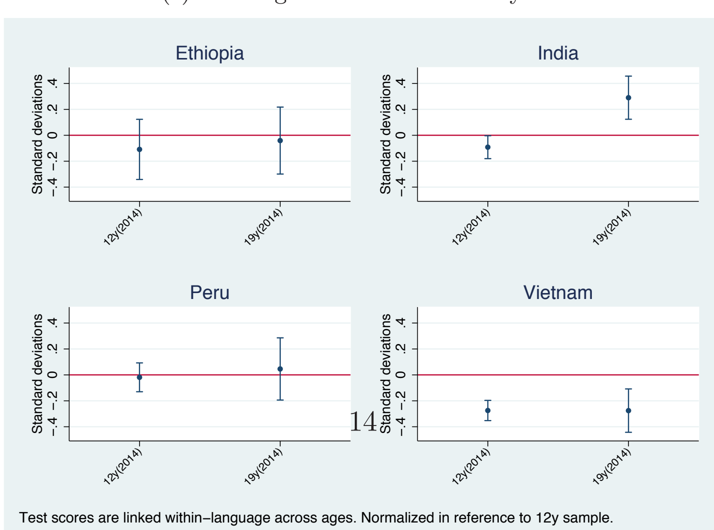
(a) Quantitative Skills from 5-19 years



(b) Receptive vocabulary from 5-15 years



(c) Reading scores at 12 and 19 years



Kolmogorov-Smirnov (K-S) tests in any country for quantitative ability, although we do reject it for Ethiopia and Peru for receptive vocabulary.[18]

At 8 years of age, similar patterns seem to hold up in mathematics; scores remain equal for boys and girls across countries, except a small gap in Peru of about 0.07 SD. In receptive vocabulary, however, significant gaps seem to open up in all countries except Ethiopia albeit with relatively modest gaps of about 0.1 SD in Peru and Vietnam but a larger gap of ˜0.15 SD in India. While these gaps favour boys in India and Peru, they are in the opposite direction in Vietnam.

At 12 years of age, there is somewhat more evidence of gender gaps in mathematics - favouring girls in Vietnam and boys in Peru and Ethiopia. In all cases, however, these gaps are modest and no more than 0.1 SD. In receptive vocabulary, we see gaps in similar direction as in mathematics in Peru about 0.15-0.2 SD in magnitude. Comparing the two cohorts at 12 years of age, to see if gender gaps have importantly shifted over the period from 2006 and 2013, we do not detect any significant differences; as is evident from Figure 2, the gender gaps at 12 in both cohorts are well within the confidence intervals of each other in all countries for both math and receptive vocabulary. On a test of reading comprehension, however, which was only administered in the fourth survey wave in 2013, we do see gender gaps in all countries except Peru; directions of the gender gap favour girls in Vietnam (consistent with other scores at this age) but favour boys in India and Ethiopia (which contrasts with results in both mathematics and receptive vocabulary in Ethiopia). The magnitude of the gap is particularly significant in Vietnam at about 0.3 SD.

At the age of 15, gender gaps are substantially larger in magnitude and almost invariably statistically significant: in math gaps of about 0.2-0.35 SD exist favouring boys in Ethiopia and India and favouring girls in Vietnam; similar differences exist also in receptive vocabulary, consistent with the gaps in math in Ethiopia and India, additionally significant favouring boys in Peru, and favouring girls but statistically insignificant in Vietnam. It appears, comparing the mean differences at 12 and 15 for the older cohort that this period from 12-15 years, i.e. the period coinciding with early adolescence and the transition out of primary schooling, coincides with a substantial development of gender gaps.

Finally, at the age of 19 years, we see gender gaps having crystallized in the expected direction (based on previous gaps) with large gaps favouring boys in Ethiopia and India and a smaller such gap also favouring boys in Peru. In Vietnam, the country where we are most likely to see gaps favouring girls, mean scores are statistically indistinguishable from each other. It appears

[18] A rejection of the null hypothesis for a two-sided K-S test cannot, of course, be used to conclude that one group first-order stochastically dominates the other — it is perfectly possible that the two non-overlapping CDFs cross each other. In those country/domain/age samples where the K-S test for the equality of means is rejected, and there is further a significant difference in means between groups, we additionally conduct Wilcoxon rank-sum tests to check that the test score distribution of the group with the higher mean (say, girls in Vietnam) first-order stochastically dominates the test score distribution of the other group. If we are unable to reject the hypothesis of first-order stochastic dominance, then we may take this as evidence that the direction of the gap is invariant to scaling decisions. Results for this are presented in Appendix 2.

thus that by early adulthood gender gaps in mathematics do favour boys wherever we find statistically significant gaps, a pattern similar to that documented in international assessments (see GMR 2012).

3.3 Understanding test score divergence

Divergence in learning by initial levels of ability

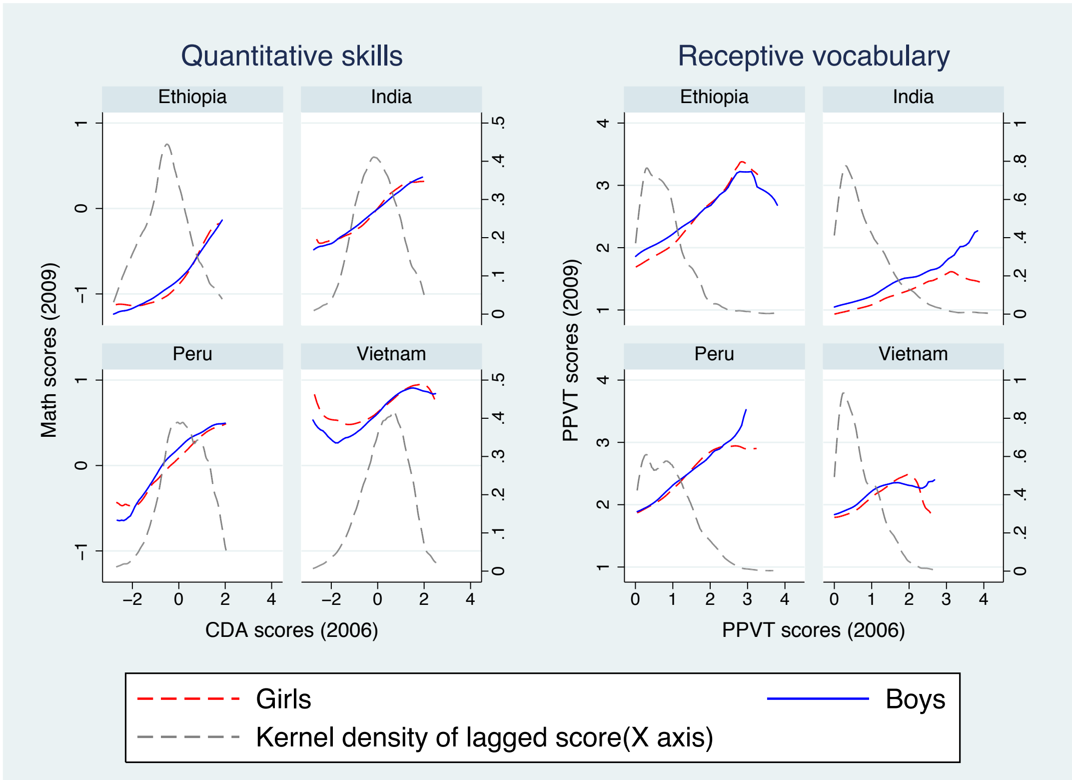
The patterns presented in Figure 2 and Table 2 show the evolution of the absolute gap in test scores between boys and girls over their age trajectories in each of the four countries. They do not however answer two critical questions. First, whether and how much the gaps seen at later ages (e.g. at 15 or 19 years) merely reflect gaps that had already arisen by 12 which are then perpetuated (and perhaps amplified) by the self-productivity of skills. Second, as importantly, is any divergence in test scores more marked at some ability levels than others? It is, for example, possible *a priori* that girls at the top end of the ability distribution continue to progress at at par with boys of similar ability but that poorly performing girls lag further behind over time.[19] Any such heterogeneity is important to capture in order to prioritize exactly which groups and age points to target in order to mitigate any gender-based divergence in test scores.

Our method for investigating these two questions is straightforward and capitalizes on the panel dimension of the data which we had previously ignored. For each age group, we present non-parametric plots relating current achievement to lagged achievement separately for boys and girls. If all of the gap at, say, 15 merely reflected the continuation of gaps at 12, then we should see the two non-parametric plots completely overlap each other for a particular country. If there is a general divergence in test scores for all boys, regardless of prior ability, then we should see the two non-parametric plots be shifted versions of each other with an intercept difference only. Finally, if the divergence does depend on initial levels, then we should see a slope difference in the two lines at different levels of prior ability.

These graphs also serve a further, important, purpose. To the extent that we see the non-parametric curve for one group (say, boys) lie above that of the other, we can be confident that our conclusions about there being divergence in test scores across genders between the two ages is robust to concerns of ordinality: since, at all levels of ability, one group makes greater progress than the other, any rank-preserving transformation of the test scores will continue to show this pattern of divergence. If the two curves also have the same slope at different levels of ability, this further implies that the magnitude of the gap will be unchanged regardless of any ordinal transformation. Our analysis of whether there is fresh divergence between particular

[19] Such a pattern need not, of course, reflect anything intrinsic to the process of skill formation: It could well reflect different choices made with respect low performing girls vs. low performing boys, for example in providing remedial investments (such as extra tutoring) to one group and not the other.

Figure 3: Divergence from 5-8 years



Lines are local polynomial smoothed lines (epanechnikov kernel and degree zero) plotted separately for boys and girls.

ages is therefore robust to the concerns of ordinality raised by Bond and Lang (2013), in contrast to much previous work around gender-based or racial gaps in achievement.

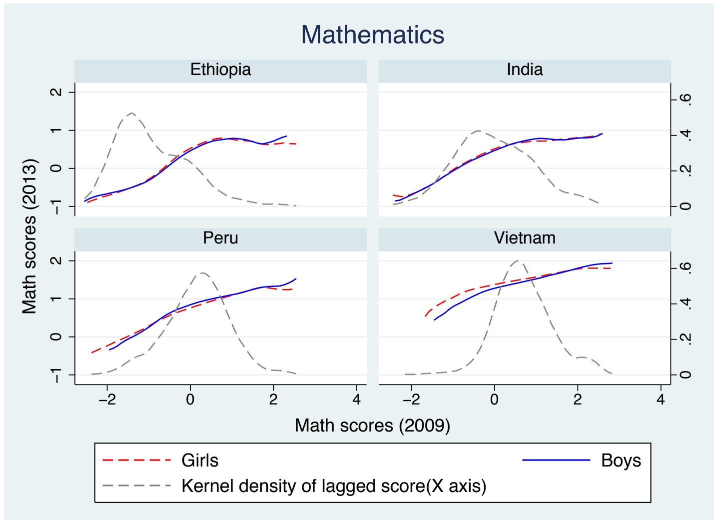
Figures 2 to 5 present these plots for divergence in mathematics test scores between 5-8 years, 8-12 years, 12-15 years and 15-19 years of age in the four countries, along with kernel densities of the lagged test score distribution in all the relevant tests adminstered at that age.

These graphs show some interesting patterns, further nuancing the picture obtained from the cross-sectional comparisons shown above. Between 5-8 years, as expected, we see a near-perfect overlap in the trajectories of boys and girls in quantitative ability. In receptive vocabulary, by contrast, we see that in India at all levels of test score at 5 boys seem to outperform girls, thus leading to at the age of 8 years the statistically significant difference in mean scores and in the distributions of achievement noted in Table 2.

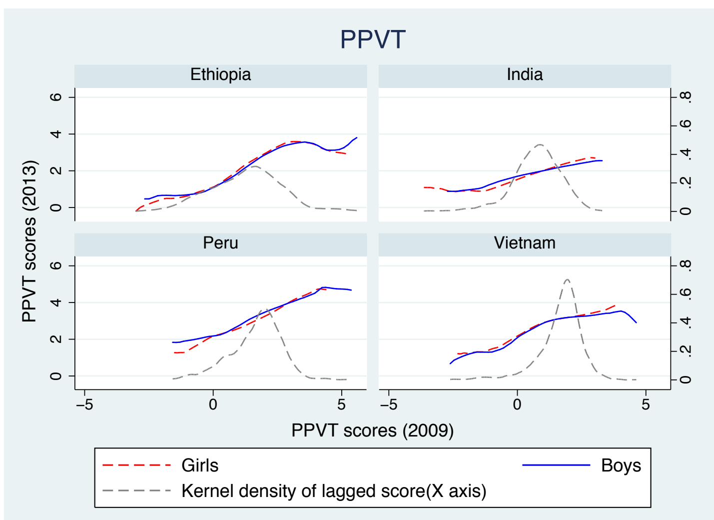
Between 8-12 years of age, again trajectories in mathematics seem to overlap near-perfectly for all countries over the bulk of the distribution of 8-year math scores. The same is true of

Figure 4: Divergence from 8-12 years

(a) Quantitative Skills



(b) Receptive vocabulary



(c) Reading scores

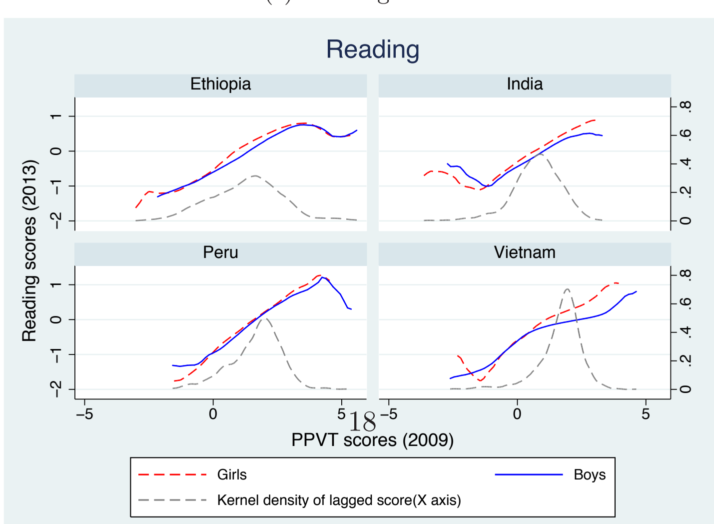
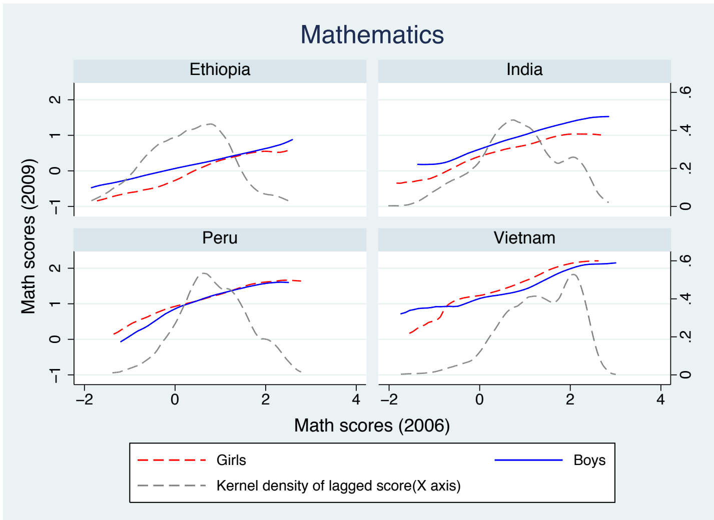
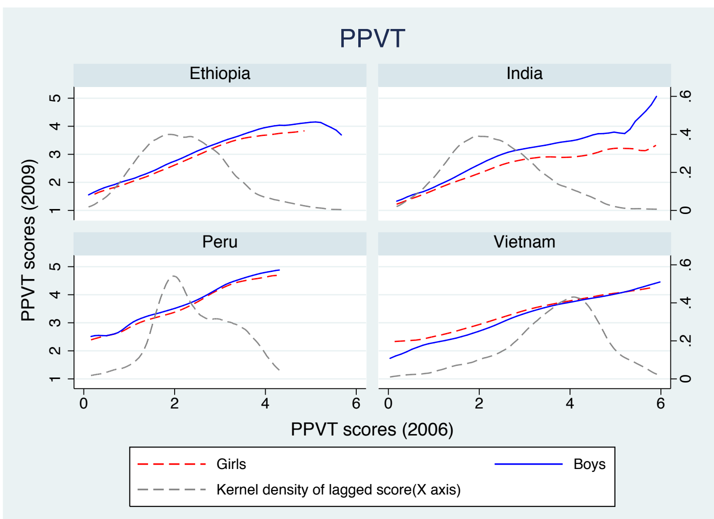


Figure 5: Divergence from 12-15 years

(a) Quantitative Skills



(b) Receptive vocabulary



(c) Cloze scores

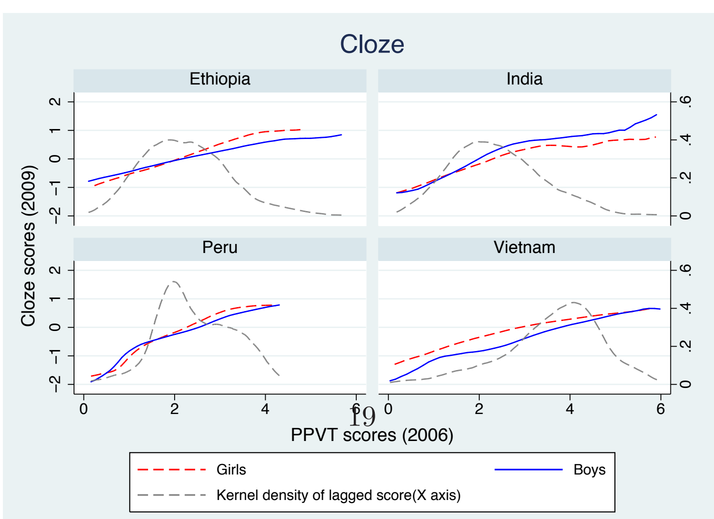
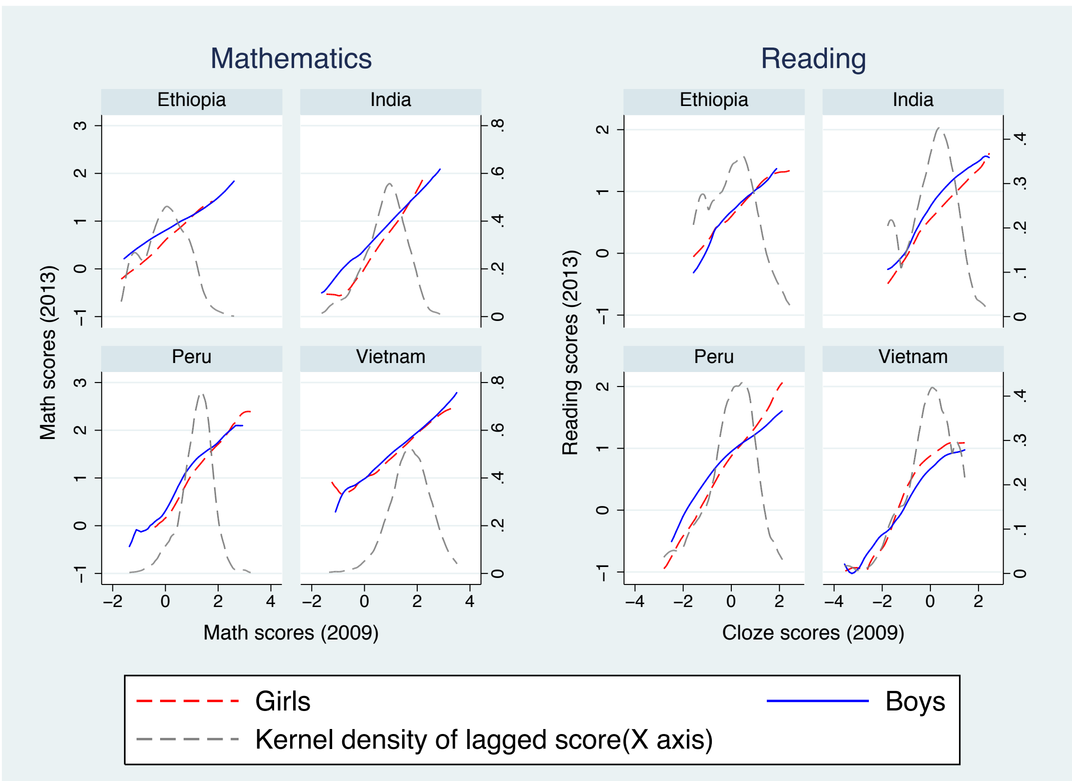


Figure 6: Divergence from 15-19 years



Lines are local polynomial smoothed lines (epanechnikov kernel and degree zero) plotted separately for boys and girls.

scores on receptive vocabulary but clear evidence of gaps emerging is evident in the reading tests. In India, for the range of variation in the lagged vocabulary scores, girls achieve a slightly higher reading score at 12 than boys. In Vietnam, this gap is even more evident and therefore naturally expressed in the larger gender gap in Table 2. However, in a pattern not evident earlier, apparently the gender gap here is most evident for initially higher-performing boys and girls, thus displaying some indication of heterogeneity by initial ability.

Between 12 to 15, as we expected based on the cross-sectional gaps presented earlier, we see strong evidence of divergence. Ethiopian boys, at all levels of ability at 12 years, score as much or more inmathematics than girls by the age of 15 thus contributing to the emergence of a gender gap. Similar patterns are evident in India, being even clearer, of greater magnitude and evident for all three learning tests administered at 15. In Vietnam, there is clear evidence of girls performing better in reading at 15 at all levels of vocabulary scores at 12; similarly in mathematics, across the bulk of variation in the sample in math scores at 12, girls outperform boys.

Finally, the patterns between 15 and 19 years reveal essentially a continuation of patterns seen between 12 and 15. In reading, we see modest divergence favouring girls in Vietnam and boys in India. In mathematics, we see divergence favouring boys in Ethiopia, India and Peru. An interesting feature of the mathematics divergence is that the slope of the curve seems considerably steeper for girls rather than boys in both Ethiopia and India, indicating that the gap is primarily concentrated at the lower end of the 12-year old mathematics distribution: girls who performed poorly at 12 in math have fallen much further behind boys who initially scored similarly, than girls who had higher scores at 12.

Decay in test scores

An interesting feature of the results in Table 2 and the figures above is that we frequently see a decline in the absolute magnitude of the gender gap from one age group to the next - e.g. in Vietnam the (pro-girl) gender gap in mathematics scores is clearly evident at 15 but has closed by 19 - even though the trajectories as displayed by the non-parametric two-way plots for boys and girls exactly overlap. Put simply, it seems that even though the rates of learning progress for boys and girls in Vietnam have been similar between 15 and 19 years, the absolute gender gap has declined into insignificance.

The explanation for this lies in the empirical regularity of high decay in test scores over time. As a number of studies have documented in different contexts, only 0.25 to 0.5 of learning (as measured by test scores) seems to persist from one year to the next (see e.g. Jacob et al. (2010), Andrabi et al. 2011) and fade-out of intervention effects is also routine. In the case of gender gaps in achievement, this is of particular importance: decay in test scores implies that there may be‘fresh’ divergence in test scores occuring even if the absolute magnitude of the gender gap measured cross-sectionally remains the same or, in some cases, even declines.[20]

This is of more than academic interest: to the extent that policy priorities are focused towards the reduction of gaps, and in particular to avoiding their exacerbation, panel-based analyses such as ours may indicate age-windows for intervention that would not be evident when only looking at the gaps cross-sectionally. Moreover, even analysis of changes in inter-group gaps is not fully informative since effectively it implicitly assumes perfect persistence of learning — given that this assumption is found routinely violated in panel data, and has important consequences for which ages we think are most necessary to focus on, this is of particular concern in any analyses intending to shed light on the evolution of skill inequalities in childhood.[21]

4 Understanding the sources of divergence from 12-15 years

Having presented the descriptive patterns in gender gaps, we focus the rest of this paper towards understanding the proximate causes of this gender-based divergence in test scores. In particular, we wish to examine if we can identify systematic gender-based differences in the factors determining learning which may serve to explain those patterns of divergence that we have identified thus far. In particular, we focus on three sets of factors that may individually or jointly explain the gaps: household characteristics and investments into child learning; the time-use of children; and differential quality of and experiences in the schools attended.[22]

In the analysis in this section, we present descriptive analysis of gender differences in the factors affecting learning for all ages. In the regression analyses, however, we choose to restrict our explanation of gender-based divergence in test scores to only one age group, namely the 15 year old sample. This choice is informed by our results from the previous section: the period between 12-15 years seems to be a key period in all countries except Peru for divergence in test scores and in all countries for the emergence of gender differences in enrolment.

[20] Put simply, whether any significant gap declines depends both on decay (which drives gaps towards zero) and the difference in trajectories (which could be in any direction). Where the differences in trajectories are in the same direction as the initial differences in levels, e.g. girls scored higher at 12 *and* had a higher trajectory between 12 and 15, whether the absolute magnitude of the gender gap increases, stays constant or declines depends on whether the additional gap caused by the higher trajectory exceeds, is the same as, or less than the decline in the gender gap to be expected naturally as a result of decay in test scores.

[21] This point is particularly important to the observation in Bond and Lang(2013b) that the pattern that a small set of covariates can ‘explain’ the Black-White gap does not mean that later periods of development do not matter in the evolution of inequality in test scores.

[22] We call these proximal factors since they may be caused in turn by more general features such as labour market opportunities in adulthood or social norms not captured here. As recent work has shown, e.g. Jensen (2012) and Munshi & Rosenzweig (2006) in India, changes in these broader economic and social factors may change patterns of differential investments in the education of boys and girls, which may reasonably be expected to affect the inequalities in human capital manifested by differences in test scores.

Table 3: Descriptives of control variables

	Ethiopia			India			Peru			Vietnam		
	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff
Household level variables												
Caregiver's educaton level												
— None	0.50	0.51	-0.00	0.68	0.69	-0.00	0.10	0.11	-0.01	0.09	0.11	-0.02
— Up to Grade 8 (Primary)	0.26	0.28	-0.02	0.19	0.21	-0.01	0.37	0.33	0.04	0.27	0.29	-0.02
— Grade 9-10 (Secondary)	0.02	0.02	0.00	0.08	0.07	0.02	0.37	0.38	-0.01	0.46	0.42	0.04
— Grades 11-12 (Higher secondary)	0.04	0.03	0.01	0.02	0.02	0.00	0.06	0.07	-0.01	0.13	0.13	-0.00
— Higher Education	0.17	0.16	0.01	0.01	0.02	-0.01	0.10	0.10	-0.01	0.06	0.06	0.00
Household size	6.36	6.34	0.02	5.02	5.08	-0.06	5.33	5.43	-0.10	4.65	4.43	0.22*
Urban	0.43	0.40	0.02	0.24	0.26	-0.02	0.76	0.78	-0.02	0.19	0.21	-0.02
Wealth index	0.35	0.35	-0.01	0.52	0.53	-0.01	0.59	0.59	-0.00	0.63	0.62	0.02
Child-specific investment variables												
Enrolled at 15 years	0.91	0.88	0.04	0.74	0.81	-0.07*	0.95	0.91	0.04	0.81	0.73	0.08**
Annual child specific expenditure on education (nominal, local currency)	130.92	192.47	-61.55	1474.66	3162.62	-1687.97***	308.99	347.62	-38.63	1935.70	1858.08	77.62
Height-for-age z-score	-1.01	-1.74	0.73***	-1.69	-1.60	-0.08	-1.59	-1.38	-0.21**	-1.40	-1.46	0.06

4.1 Household investments into education

An obvious area to first look for explanations lies in the household characteristics and resource allocation between boys and girls.[23] This is an especially pertinent area to look at since in many domains of human capital in childhood, we know that differences in outcomes result directly from household level choices regarding investment into children.[24]

We look first at a relatively parsimonious and standard set of household characteristics and investments. For household characteristics, we use three main variables: caregiver's education, a wealth index, and household size. With regard to investments and child-specific characteristics, we use child-specific information on enrolment, child-specific expenditures on education and children's nutritional status, as summarized by WHO height-for-age z-scores which encapsulate endowments and longer-term investments into health. Differences by gender with regard to all household characteristics and investments, with the exception of enrolment which was already summarized in Table 1, are provided in Table 3.

As may be expected if gender is near-randomly distributed, there is not much evidence of differences in the household characteristics of boys and girls in any of our countries. The allocation of investments, however, reveals significant differences. As we had noted while

[23]In general, treating the sex of a child as a random event, we would not expect to find systematic differences between the household characteristics of boys and girls. However, if some contexts have selection in sex of the child, for example through sex-selective abortions or selective stopping rules for fertility arising from son-preference or gender differences in infant mortality, then systematic differences in household characteristics may still exist. Given that this possibility cannot be ruled out in all our contexts, with previous evidence suggesting the existence of such channels in India (see e.g. Bhalotra xxx, Dasgupta xxx), levels differences in household characteristics still merit consideration as a possible channel for gender differences in learning.

[24]In the Indian setting for example, ...[add stuff]

discussing Table 1, enrolment at 15 and 19 years displayed significant gender differences in both India and Vietnam. Education expenditures are lower for girls than boys in all countries except Vietnam at 15, although only significant in India where the average annual expenditure on boys' education is double the amount spent on girls. Finally, in nutrition, we do not see many clear patterns of gender gaps in the height-for-age z-scores of children in the four countries at 15 although there is some evidence of a modest gender gap favouring girls in Ethiopia and boys in Peru.

The method for the decomposition of gaps is straightforward as in the following specifications:

$$
\begin{aligned}
Y_{ia} &= \alpha + \beta_1.male_i && (1)\\
&\quad +\beta_2.Y_{i,a-1} && (2)\\
&\quad + + \beta_3.X_i + \beta_4.enrol_{i,a} && (3)\\
&\quad +\beta_5.EdExp_{ia} && (4)\\
&\quad +\beta_6.HAZ_{ia} + \epsilon_{ia} && (5)
\end{aligned}
$$

where Y_{ia} is the current test score for child i at age a, male is a dummy variable (1=male), $Y_{i,a-1}$ is the lagged test score from the previous age group, enrol is a dummy variable denoting current enrolment, $EdExp$ denotes household spending on the particular child's education, and HAZ_{ia} is the height-for-age z-score; $\epsilon_i a$ is a disturbance term. Each regression is run within-country thus allowing the coefficients to vary across countries.

The sequential inclusion of inputs in equations (1)-(5) naturally changes the intepretation of β_1, our main coefficient of interest. Equation (1) merely shows mean differences by gender as in Table 2; in Equation 2, it shows the divergence across gender in a particular age, conditional on past test score, a linear analogue of Figure 3; Equation (3) additionally controls for enrolment and household characteristics, seeking to investigate to what extent the gender-based differences in enrolment in Table 1 might account for the gender gaps in Table 2; Equation (4) additionally controls for differences in the educational expenditure for the individual child i; finally Equation (5) adds the anthropometric z-scores to further see if systematic differences in nutrition between boys and girls can explain gender gaps in student achievement.

It is worthwhile discussing the interpretation of estimates from the above exercise. As Fortin et al. (2011) note, drawing clear parallels between the decomposition literature and the causal treatment effects literature, in order to make counterfactual statements – for example, how much would equalizing enrolment for boys and girls close gender gaps in test scores – decomposition exercises also rely on a ignorability (conditional exogeneity) condition. Is such a condition justifiable in this setting, for example in interpreting the coefficients on enrolment, expenditures or time use?

The inclusion of lagged measures of achievement essentially converts Specifications (3)-(5) into a dynamic OLS value-added model (VAM). These specifications can be derived from a general cumulative effects model of education production (see e.g. Andrabi et. al. 2011, Todd and Wolpin 2003) where essentially the past score provides a summary measure of all past investments and individual-specific heterogeneity, conditional on which ignorability may be assumed to hold.[25] While it is possible that estimates from VAMs are still biased due to measurement error and unobserved heterogeneity, in practice the extent of bias seems to low across a range of applications. In particular, estimates from these models have proven to be unbiased in comparison to experimental estimates (e.g. Deming, 2014; Deming et al., 2014; Kane et al., 2013; Angrist et al., 2013; Singh, 2015) and to rigorous quasi-experimental estimates (e.g. Chetty et al. 2014; Andrabi et al. 2011; Singh 2014).

In an ideal setting, we would have experimental variation in each element of the input vector. However, although various randomized trials do vary elements of the input vector, it is likely impracticable to have experimental variation simultaneously in *all* inputs. Similarly, it is also very difficult to identify as many valid instruments as the number of inputs. Given these binding feasibility constraints, and the fact that no such data exist in a comparative setting to our knowledge, a VAM framework is, perhaps, the best available method to investigate the sources of gender gaps in learning and, in our judgment, a significant advance over cross-sectional decompositions which are the most that are possible using most comparative sources of achievement data such as PISA and TIMSS.[26]

Note that, although parsimonious, the list of investments do have good summary measures for household-based investments and therefore, *a priori*, we expect that they should be able to account for a substantial share of the variance in learning outcomes. The lagged test scores should, in a cumulative effects VAM, proxy for the full range of past investments and ability. Educational expenditures could summarize a range of different investments into education including e.g. the type of school, extra tutoring, extra support for buying books and school materials etc. Similarly, the height-for-age z-scores should be able to proxy for early childhood investments in health and nutrition.

[25] Note that the ignorability condition does not require all assumptions of the structural cumulative effects model to hold. It requires merely that, conditional on lagged achievement, the inputs are uncorrelated with the error term, a weaker condition. This is akin to the justification behind propensity score matching methods and indeed an essentially similar specification is justified as such, and shown to be unbiased in comparison to lottery-based estimates, by Angrist et. al. (2013) in their study of Charter Schools in the US.

[26] A further possible source of confusion is whether coefficients in these models should be interpreted as technology parameters or treatment effects, a judgment that depends on the full list of variables being controlled for (see Todd and Wolpin, 2003; Singh 2015 for a discussion). In this paper, we side-step this issue since we are not interested in the input coefficients *per se* but rather in whether these inputs, in unison, can explain the patterns in gender gaps that we see in the data.

Results from the decomposition exercise are presented in Table 4-6 for mathematics, receptive vocabulary and the cloze test respectively. The major result, across the different tests and countries, is that although the various factors affecting learning do have expected signs, and for the most important hypothesised factors are statistically significant, they jointly explain only a fraction of the opening up of learning gaps in this age period. At best, in the Indian case, they can explain about half of the gender gap in the cloze test and about a third in mathematics and receptive vocabulary. In Vietnam, the coefficient on the male dummy does decline by about a third in both math and receptive vocabulary but in the cloze test is practically unchanged across specifications. In Ethiopia, it appears that the range of controls have jointly have near-zero explanatory power regarding the gender gap.

Why is this set of relatively rich summary measures affecting achievement still relatively ineffective at explaining gender based divergence in teenage years? To an extent, the inability of past achievement and various household characteristics to explain the gender gaps may be expected based on previous results: gender differences in these characteristics are either insignificant or small in magnitude, thus already suggesting that the explanation is unlikely to be found in the difference in levels of these factors between sexes. That we can explain the gender gaps in achievement substantially in India is not because the factors are jointly more predictive in India - indeed the R-square is generally similar across countries - but because it is the only sample wherein lagged achievement, enrolment and child-specific educational expenditures all display statistically significant gender bias in the same direction as the gender gap in learning. Put differently, it is only in the Indian sample that we see clear differences in the investments underlying achievement production, which are directly-measured in our data, and therefore it is in India that we can explain a substantial portion of the gender gaps in achievement by 15 years of age. Nevertheless, this still leaves open the question about where the gaps in achievement between boys and girls could be arising from, given that they seem not to be arising from the most commonly considered inputs into learning.

4.2 Time use of adolescents

One possibility that perhaps could explain the emergence of the gender gap between 12 to 15 years relates to the time use of children. Specifically, it is possible that, even if boys and girls had similar achievement at the age of 12 years, and similar rates of progression through school, differences may open up in how they allocate their time between studying and other uses such as household chores, paid work or leisure. The time children expend in studying, proxying for their effort, may systematically change for boys and girls during this period of adolescence and may explain differences in subsequent achievement.

Table 4: Basic decomposition results: Mathematics

			Ethiopia					India					Peru					Vietnam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
VARIABLES										*Dependent var: Math scores at 15 years*										
Male	0.247***	0.206***	0.210***	0.207***	0.223***	0.334***	0.287***	0.274***	0.246***	0.241***	-0.00284	-0.0503	-0.0296	-0.0336	-0.0421	-0.187***	-0.156***	-0.132***	-0.132***	-0.128**
	(0.0555)	(0.0594)	(0.0534)	(0.0530)	(0.0583)	(0.0599)	(0.0622)	(0.0552)	(0.0577)	(0.0573)	(0.0553)	(0.0526)	(0.0475)	(0.0472)	(0.0471)	(0.0377)	(0.0391)	(0.0453)	(0.0452)	(0.0476)
Currently in school			0.300***	0.297***	0.287***			0.229***	0.204***	0.209***			0.370***	0.360***	0.361***			0.346***	0.319***	0.318***
			(0.0770)	(0.0763)	(0.0740)			(0.0636)	(0.0615)	(0.0622)			(0.0650)	(0.0627)	(0.0636)			(0.0580)	(0.0627)	(0.0615)
Expenditure on child education				5.33e-05**	0.000233***				1.66e-05***	1.57e-05***				0.000109**	0.000105**				2.90e-05*	2.80e-05*
				(2.51e-05)	(5.89e-05)				(4.61e-06)	(4.36e-06)				(3.88e-05)	(3.98e-05)				(1.48e-05)	(1.48e-05)
Height-for-age z-score					0.0258					0.0569**					0.0401*					0.0830***
					(0.0246)					(0.0208)					(0.0218)					(0.0243)
Lagged math score		0.498***	0.453***	0.448***	0.437***		0.439***	0.371***	0.365***	0.360***		0.484***	0.410***	0.408***	0.401***		0.503***	0.383***	0.377***	0.361***
		(0.0296)	(0.0291)	(0.0292)	(0.0308)		(0.0450)	(0.0476)	(0.0475)	(0.0464)		(0.0294)	(0.0340)	(0.0317)	(0.0311)		(0.0438)	(0.0514)	(0.0505)	(0.0512)
Constant	-0.114	-0.227***	-0.540***	-0.536***	-0.500***	0.634***	0.270***	-0.203*	-0.122	-0.00668	1.184***	0.798***	0.349**	0.376**	0.479***	1.804***	1.126***	0.758***	0.798***	0.970***
	(0.0867)	(0.0616)	(0.131)	(0.131)	(0.127)	(0.0707)	(0.0730)	(0.117)	(0.126)	(0.142)	(0.0593)	(0.0407)	(0.166)	(0.165)	(0.150)	(0.0872)	(0.0892)	(0.181)	(0.178)	(0.175)
Observations	931	894	890	890	880	904	895	883	883	879	661	658	656	656	654	951	943	903	903	901
R-squared	0.023	0.311	0.361	0.363	0.372	0.045	0.291	0.358	0.366	0.370	0.000	0.345	0.405	0.413	0.415	0.014	0.269	0.342	0.348	0.357

Table 5: Basic decomposition results: Receptive vocabulary

			Ethiopia					India					Peru					Vietnam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
VARIABLES										*Dependent var: Vocabulary scores at 15 years*										
Male	0.229	0.124	0.185***	0.181***	0.219***	0.412***	0.349***	0.336***	0.324***	0.319***	0.145	0.0695	0.0894*	0.0874*	0.0799	-0.0475	-0.0840*	-0.0405	-0.0402	-0.0289
	(0.131)	(0.0815)	(0.0524)	(0.0508)	(0.0382)	(0.0775)	(0.0754)	(0.0781)	(0.0820)	(0.0793)	(0.0877)	(0.0534)	(0.0497)	(0.0496)	(0.0500)	(0.0601)	(0.0484)	(0.0492)	(0.0487)	(0.0526)
Currently in school			0.114	0.112	0.113			0.354***	0.344***	0.352***			0.188*	0.183*	0.184*			0.277***	0.260**	0.257**
			(0.197)	(0.198)	(0.195)			(0.0708)	(0.0722)	(0.0706)			(0.107)	(0.106)	(0.105)			(0.0875)	(0.0990)	(0.0978)
Expenditure on child education				3.02e-05***	0.000115***				7.79e-06	5.96e-06				6.28e-05	5.92e-05				1.72e-05	1.53e-05
				(8.37e-06)	(3.59e-05)				(7.72e-06)	(7.83e-06)				(4.32e-05)	(4.26e-05)				(2.70e-05)	(2.65e-05)
Height-for-age z-score					0.0572*					0.0922***					0.0436					0.147***
					(0.0286)					(0.0253)					(0.0358)					(0.0496)
Lagged vocabulary score		0.766***	0.392***	0.385***	0.358***		0.577***	0.437***	0.434***	0.426***		0.810***	0.669***	0.664***	0.656***		0.510***	0.329***	0.327***	0.302***
		(0.114)	(0.0953)	(0.0981)	(0.0960)		(0.0828)	(0.0819)	(0.0821)	(0.0798)		(0.0249)	(0.0327)	(0.0327)	(0.0343)		(0.0510)	(0.0477)	(0.0466)	(0.0474)
Constant	2.599***	1.034***	0.947**	0.960***	1.129***	2.353***	1.045***	0.474***	0.517***	0.721***	3.707***	1.766***	1.363***	1.386***	1.505***	3.537***	1.743***	1.033***	1.057***	1.393***
	(0.279)	(0.299)	(0.317)	(0.319)	(0.301)	(0.0925)	(0.153)	(0.154)	(0.180)	(0.176)	(0.133)	(0.0774)	(0.131)	(0.131)	(0.125)	(0.168)	(0.133)	(0.229)	(0.221)	(0.221)
Observations	462	413	406	406	399	895	889	813	813	809	645	635	629	629	628	965	960	902	902	900
R-squared	0.010	0.431	0.560	0.561	0.560	0.045	0.416	0.461	0.462	0.472	0.006	0.529	0.566	0.568	0.569	0.000	0.352	0.394	0.395	0.410

Table 6: Basic decomposition results: Cloze language test

			Ethiopia					India					Peru					Vietnam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
VARIABLES										*Dependent var: Cloze scores at 15 years*										
Male	0.0132	-0.103	-0.0509	-0.0540	-0.0540	0.233***	0.196**	0.166**	0.121*	0.121*	-0.0409	-0.121*	-0.102	-0.117*	-0.117*	-0.238***	-0.258***	-0.229***	-0.223***	-0.223***
	(0.173)	(0.160)	(0.147)	(0.177)	(0.177)	(0.0805)	(0.0698)	(0.0624)	(0.0594)	(0.0594)	(0.101)	(0.0654)	(0.0657)	(0.0668)	(0.0668)	(0.0597)	(0.0709)	(0.0659)	(0.0619)	(0.0619)
Currently in school			0.407***	0.432***	0.432***			0.550***	0.515***	0.515***			0.295**	0.296**	0.296**			0.406***	0.386***	0.386***
			(0.0736)	(0.0671)	(0.0671)			(0.0848)	(0.0784)	(0.0784)			(0.108)	(0.111)	(0.111)			(0.0639)	(0.0641)	(0.0641)
Expenditure on child education				7.27e-05	7.27e-05				2.55e-05***	2.55e-05***				-8.03e-05**	-8.03e-05**				2.24e-05	2.24e-05
				(5.52e-05)	(5.52e-05)				(6.91e-06)	(6.91e-06)				(3.62e-05)	(3.62e-05)				(1.79e-05)	(1.79e-05)
Height-for-age z-score				0.00699	0.00699				0.0605**	0.0605**				0.0903*	0.0903*				0.0713*	0.0713*
				(0.0515)	(0.0515)				(0.0269)	(0.0269)				(0.0465)	(0.0465)				(0.0368)	(0.0368)
Lagged vocabulary score		0.608***	0.386***	0.379***	0.379***		0.550***	0.397***	0.383***	0.383***		0.758***	0.580***	0.569***	0.569***		0.413***	0.258***	0.241***	0.241***
		(0.0801)	(0.0707)	(0.0754)	(0.0754)		(0.0746)	(0.0764)	(0.0724)	(0.0724)		(0.0412)	(0.0529)	(0.0513)	(0.0513)		(0.0343)	(0.0463)	(0.0481)	(0.0481)
Constant	-0.00645	-1.285***	-1.577***	-1.579***	-1.579***	-0.114	-1.371***	-1.898***	-1.626***	-1.626***	0.0205	-1.802***	-2.279***	-2.053***	-2.053***	0.117	-1.364***	-1.272***	-1.080***	-1.080***
	(0.245)	(0.210)	(0.269)	(0.261)	(0.261)	(0.0716)	(0.127)	(0.144)	(0.153)	(0.153)	(0.136)	(0.123)	(0.161)	(0.183)	(0.183)	(0.0924)	(0.142)	(0.231)	(0.215)	(0.215)
Observations	462	417	410	403	403	859	854	786	782	782	660	650	645	643	643	952	948	899	897	897
R-squared	0.000	0.331	0.381	0.377	0.377	0.014	0.324	0.354	0.369	0.369	0.000	0.424	0.471	0.478	0.478	0.014	0.267	0.282	0.289	0.289

Such systematic gender differences may arise for several reasons. For example, gender-specific demands on tasks apart from education may rise more for one group than the other e.g. if girls are expected to contribute more at this age to household chores or if boys are expected to contribute financially to the household requiring dedication of time to paid work. They could arise if social norms are reinforced and internalized which encourage effort by one sex and not the other. Finally, they could arise as adolescents become aware of any gender-based differential in the return to human capital in adulthood and therefore adjust their own effort accordingly (e.g. if better academic results are likely to result in a better job for men but not women). These possibilities are not mutually exclusive, and we will not be able to distinguish between them empirically in our data, but it seems worthwhile to investigate any proximate differences in the time allocation of children to investigate whether (a) time allocation patterns differ systematically between boys and girls, (b) is this particularly a difference that opens up or widens in the 12-15 years period, (c) does the allocation of time across different purposes predict achievement in value-added models and (d) does any gender-based difference in time allocation allow us to explain better the widening of gender gaps in this age group?

The Young Lives data collect, for each of the age groups being studied in this paper, their time allocation across different uses on a 'typical' day. Such information on time allocation is relatively rare in developing country contexts and we are aware of few attempts to study the effects on learning measures.[27] As can be seen in Table 7, there is considerable indication of systematic gender differences in time allocation in all countries: in particular, girls spend more time on average on domestic tasks and chores while boys often spend more time working on the family farm or outside the household and that these work-related activities become increasingly important after 12 years of age as we would expect; the effect of these differential time allocation patterns on learning gaps remains an open question. Looking at direct time inputs into learning, we see that whereas there are few differences in the time spent at school at the age of 12 years, there is a small difference already evident in the time spent studying after school by boys and girls and that this difference widens by the age of 15 years; in India, Ethiopia and Vietnam, this is also in the same direction as the gender gaps in learning.[28]

It is relevant to note that differences in Table 7 do not correct for differential enrolment rates by sex, which is biased in favour of boys in India and girls in Ethiopia, Peru and Vietnam at 15 years of age. In Peru, India and Vietnam, the significantly different enrolment at 15 likely accounts for at least part of the significant difference by sex in the time spent at school or studying after school: to the extent that this differential allocation is already captured by our controlling for

[27] Although it should be noted that time use information has previously been used to explain achievement in the Young Lives data, notably by Singh (2015) in India and Singh (2014), Keane and Krutikova (2015) and Borga (2015) in a cross-country setting.

[28] The magnitude of the difference does not seem too large and is typically about a quarter of an hour a day going up to about 36 minutes per day extra studying for girls in Vietnam at 15.

Table 7: Gender differences in time allocation at different ages

	Ethiopia			India			Peru			Vietnam		
	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff	Female	Male	Diff
8 years (2009)												
Caring for others	1.07	0.61	0.46***	0.24	0.18	0.07**	0.49	0.47	0.02	0.32	0.17	0.14***
Domestic tasks and chores	2.06	1.29	0.77***	0.44	0.24	0.20***	0.88	0.86	0.02	0.63	0.47	0.15***
Tasks on domestic farm/business	0.80	2.13	-1.33***	0.01	0.02	-0.01	0.23	0.27	-0.04	0.07	0.16	-0.09**
Work outside household	0.00	0.03	-0.03*	0.01	0.00	0.01	0.00	0.01	-0.00	0.00	0.00	0.00
At school	4.90	4.90	0.01	7.62	7.71	-0.10	5.99	5.97	0.02	4.96	4.99	-0.02
Studying	0.98	1.00	-0.02	1.90	1.77	0.12*	1.90	1.82	0.09*	2.80	2.73	0.07
Play	4.48	4.41	0.08	4.67	4.88	-0.21**	4.05	4.30	-0.25**	5.47	5.75	-0.28***
Sleep	9.70	9.64	0.06	9.11	9.17	-0.06	9.67	9.61	0.07	9.72	9.71	0.01
12 years (2014)												
Caring for others	0.83	0.47	0.36***	0.19	0.09	0.10***	0.86	0.79	0.06	0.48	0.37	0.11**
Domestic tasks and chores	2.33	1.36	0.97***	1.03	0.71	0.33***	1.27	1.16	0.11**	1.28	1.00	0.28***
Tasks on domestic farm/business	0.84	2.21	-1.37***	0.12	0.17	-0.05	0.44	0.60	-0.16***	0.37	0.46	-0.09*
Work outside household	0.05	0.09	-0.04	0.07	0.05	0.01	0.06	0.08	-0.02	0.02	0.04	-0.01
At school	5.75	5.52	0.23**	7.92	8.06	-0.15	6.09	5.97	0.12**	5.44	5.32	0.12
Studying	1.55	1.45	0.11*	1.95	1.89	0.06	1.88	1.87	0.01	2.74	2.53	0.21**
Play	3.34	3.52	-0.18*	3.77	4.06	-0.29***	3.67	3.64	0.03	4.74	5.14	-0.40***
Sleep	9.29	9.38	-0.08	8.96	8.97	-0.01	9.50	9.56	-0.06	8.92	9.12	-0.21***
12 years (2006)												
Caring for others	0.78	0.45	0.33***	0.27	0.10	0.17***	0.88	0.60	0.28***	0.39	0.25	0.13**
Domestic tasks and chores	2.86	1.65	1.21***	1.24	0.55	0.69***	1.16	0.98	0.19***	1.41	1.01	0.39***
Tasks on domestic farm/business	0.88	2.04	-1.16***	0.20	0.33	-0.14	0.32	0.37	-0.05	0.59	0.70	-0.12
Work outside household	0.13	0.17	-0.04	0.40	0.37	0.03	0.03	0.15	-0.12**	0.07	0.02	0.06
At school	5.54	5.35	0.19	6.08	6.12	-0.04	5.64	5.47	0.18*	4.43	4.37	0.05
Studying	1.72	1.75	-0.03	1.83	2.02	-0.19*	2.08	1.82	0.26***	3.03	2.69	0.34***
Play	2.56	3.03	-0.47***	3.79	4.31	-0.53***	2.16	2.32	-0.17	5.47	5.99	-0.52***
Sleep	9.04	9.04	-0.00	9.04	9.04	-0.00	9.29	9.29	0.00	8.62	8.92	-0.30***
15 years (2009)												
Caring for others	0.92	0.48	0.44***	0.45	0.10	0.35***	0.82	0.67	0.15	0.23	0.11	0.12**
Domestic tasks and chores	3.48	1.76	1.72***	2.04	0.84	1.20***	1.70	1.18	0.52***	1.62	1.32	0.30***
Tasks on domestic farm/business	0.43	2.23	-1.79***	0.45	0.54	-0.09	0.69	0.66	0.03	0.93	1.25	-0.32*
Work outside household	0.32	0.51	-0.19	1.02	1.05	-0.03	0.23	0.58	-0.36**	0.48	0.59	-0.11
At school	5.74	5.29	0.46**	6.02	6.78	-0.76***	6.07	5.76	0.31*	4.32	3.93	0.38*
Studying	1.82	1.90	-0.08	1.88	2.13	-0.25*	2.27	1.94	0.33***	3.30	2.70	0.59***
Play	2.63	3.19	-0.56***	3.87	4.26	-0.39**	3.09	3.38	-0.29**	4.53	5.10	-0.57***
Sleep	8.66	8.65	0.01	8.26	8.30	-0.04	8.86	8.94	-0.08	8.46	8.91	-0.45***
19 years (2014)												
Caring for others	0.97	0.26	0.71***	1.31	0.17	1.14***	2.12	0.43	1.68***	0.81	0.22	0.59***
Domestic tasks and chores	3.19	1.22	1.97***	2.65	1.11	1.55***	2.05	1.02	1.03***	1.82	1.08	0.74***
Tasks on domestic farm/business	0.88	2.46	-1.58***	0.96	1.24	-0.27	0.66	0.63	0.02	1.08	1.57	-0.50**
Work outside household	1.19	2.11	-0.92***	1.31	2.89	-1.58***	2.08	3.75	-1.67***	2.45	3.12	-0.67*
At school	3.77	3.42	0.35	3.19	4.24	-1.05***	3.34	3.84	-0.51	3.01	2.47	0.54*
Studying	1.65	1.58	0.07	1.13	1.24	-0.11	1.49	1.47	0.02	1.32	1.09	0.24*
Play	3.75	4.54	-0.79***	5.07	5.00	0.07	3.43	3.79	-0.36*	5.23	6.12	-0.89***
Sleep	8.61	8.42	0.19*	8.37	8.11	0.26***	8.33	8.15	0.18	8.27	8.28	-0.01

enrolment in the previous section, we do not expect the differential allocation to further explain learning gaps.

To investigate the additional role of differential time allocation to the emergence of gender gaps between 12 and 15, over and above the controls examined in the previous section including enrolment, we add a full vector of time use categories to Equation (5), the most extensive specification shown in the previous sub-section. In adding the full vector of time use, we follow the lead of Fiorini & Keane (2014) who study similarly the role of time allocation in the production of cognitive skills in Australia.[29]

Results from this exercise in each of the countries is presented in Table 8 for all three tests administered to 15-year-old individuals in the survey. The central message from this table is that the information on time allocation of individuals adds little additional to our ability to explain the emergence of gender differences in learning at this age. In no country do we find significant evidence of a decline in the absolute size of the coefficient on the male dummy variable. This is not to say that the information in this vector of time of time use is irrelevant - as may be seen, both time spent in school and studying have positive coefficients and are frequently statistically significant - but rather that the information was previously already proxied by the controls we had included. In particular, whereas enrolment was consistently positive in sign and (with only the exception of receptive vocabulary in Ethiopia) always statistically significant, the inclusion of the time use categories reduces this variable to statistical insignificance in all regressions indicating that the relevant variation from time use was already proxied by the enrolment variable. Whereas time use differences may have been promising *a priori* as potential sources of divergence, their explanatory power in this case appears minimal.

4.3 The quality of schooling and experience within schools

Our focus thus far has been on trying to explain the widening of gender gaps between 12 and 15 years of age by looking at household-based sources of divergence - household based investments in children's education and nutrition, their prior achievement, and individual time use. Schooling, while clearly important to understanding a divergence in learning skills between boys and girls, has been accounted for only indirectly - through enrolment, child-specific enrolment expenditures, time spent in school and, possibly, time studying after school (which may be considered to be jointly determined by individuals and schools).

It is possible however that these indirect measures of schooling do not capture much variation in school quality that could plausibly lead to widening gender gaps in learning. The period

[29]Since the number of hours in a day add up to 24, this requires the omission of one category of time use. Here, we choose to omit the number of hours that were spent sleeping. The coefficient on each category of time use therefore should be interpreted as the increment in the productivity of an hour spent in any particular category over an hour spent sleeping.

Table 8: Do time allocation patterns explain learning divergence across gender?

	Ethiopia			India			Peru			Vietnam		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
VARIABLES	Math	Vocabulary	Cloze	Math	Vocabulary	Cloze	Math	Vocabulary	Cloze	Math	Vocabulary	Cloze
male	0.202***	0.309***	-0.170*	0.237***	0.330***	0.177***	-0.0134	0.120**	-0.104*	-0.130***	-0.00636	-0.181***
	(0.0557)	(0.0874)	(0.0942)	(0.0493)	(0.0533)	(0.0579)	(0.0394)	(0.0546)	(0.0594)	(0.0442)	(0.0579)	(0.0548)
Hours per day spent:												
— in caring for hh members	0.0162	-0.0492	-0.0644	-0.0248	0.0367	0.0204	-0.000556	0.00174	0.00695	-0.0519	-0.0102	-0.000505
	(0.0349)	(0.0564)	(0.0618)	(0.0468)	(0.0412)	(0.0430)	(0.0180)	(0.0207)	(0.0241)	(0.0328)	(0.0425)	(0.0443)
—in hh chores	-0.000571	0.0626	-0.0436	0.0782**	0.101***	0.0777**	0.0252	0.0406	0.0146	-0.0139	0.0539	0.0764**
	(0.0272)	(0.0438)	(0.0487)	(0.0321)	(0.0374)	(0.0388)	(0.0194)	(0.0285)	(0.0262)	(0.0271)	(0.0403)	(0.0388)
—in domestic tasks - farming, business	0.0121	-0.0564	0.0287	0.0505	0.0329	-0.0360	0.0206	-0.00682	0.000714	0.00679	-0.0148	-0.00593
	(0.0267)	(0.0398)	(0.0467)	(0.0315)	(0.0351)	(0.0378)	(0.0166)	(0.0252)	(0.0255)	(0.0216)	(0.0290)	(0.0307)
—in paid activity	-7.98e-05	-0.0450	-0.0106	0.0405	0.0603*	0.00907	0.0326**	0.0243	0.0105	-0.0110	-0.0188	-0.0151
	(0.0267)	(0.0473)	(0.0522)	(0.0290)	(0.0322)	(0.0364)	(0.0155)	(0.0199)	(0.0243)	(0.0185)	(0.0255)	(0.0274)
—at school	0.0286	0.0304	0.132***	0.0834***	0.108***	0.0704**	0.0383**	0.0281	0.0279	0.0331	0.0253	0.0875**
	(0.0273)	(0.0409)	(0.0413)	(0.0284)	(0.0329)	(0.0357)	(0.0163)	(0.0229)	(0.0300)	(0.0267)	(0.0406)	(0.0379)
—studying outside school	0.119***	0.107**	0.0887*	0.0992***	0.134***	0.0916***	0.0673***	0.108***	0.0621*	0.00273	0.00280	0.0179
	(0.0272)	(0.0481)	(0.0504)	(0.0247)	(0.0315)	(0.0340)	(0.0206)	(0.0299)	(0.0317)	(0.0207)	(0.0277)	(0.0254)
—leisure activities	-0.00775	0.0126	0.0319	0.0539**	0.0734**	0.00153	0.0180	0.0497***	0.0340	-0.0195	-0.00555	-0.0181
	(0.0245)	(0.0380)	(0.0419)	(0.0261)	(0.0290)	(0.0329)	(0.0153)	(0.0186)	(0.0247)	(0.0188)	(0.0259)	(0.0249)
Lagged math score	0.410***			0.346***			0.391***			0.354***		
	(0.0285)			(0.0259)			(0.0305)			(0.0329)		
Lagged vocabulary		0.357***	0.328***		0.401***	0.370***		0.641***	0.567***		0.299***	0.228***
		(0.0565)	(0.0563)		(0.0294)	(0.0304)		(0.0434)	(0.0440)		(0.0326)	(0.0301)
Constant	-0.546	0.925	-1.584**	-0.826*	-0.330	-1.877***	0.180	1.194***	-2.197***	1.111***	1.399***	-1.097**
	(0.362)	(0.602)	(0.648)	(0.432)	(0.472)	(0.526)	(0.212)	(0.265)	(0.338)	(0.322)	(0.413)	(0.440)
Observations	880	399	403	875	805	779	653	627	642	895	894	891
R-squared	0.397	0.587	0.434	0.386	0.493	0.395	0.429	0.583	0.486	0.364	0.423	0.306

Table 9: Does sorting across schools account for divergence in learning?

VARIABLES	(1) Math	(2) Ethiopia Vocabulary	(3) Cloze	(4) Math	(5) India Vocabulary	(6) Cloze	(7) Math	(8) Vietnam Vocabulary	(9) Cloze
Male	0.230***	0.226***	0.0212	0.151*	0.206**	-0.0890	-0.0568	0.00538	-0.184***
	(0.0505)	(0.0840)	(0.0914)	(0.0874)	(0.0803)	(0.104)	(0.0512)	(0.0628)	(0.0671)
Lagged math score	0.386***			0.343***			0.298***		
	(0.0325)			(0.0442)			(0.0400)		
Lagged vocabulary score		0.373***	0.373***		0.532***	0.449***		0.269***	0.212***
		(0.0680)	(0.0623)		(0.0563)	(0.0605)		(0.0363)	(0.0386)
Constant	-0.211*	1.443***	-0.929***	0.338	1.085***	-1.042***	1.271***	1.671***	-0.826***
	(0.116)	(0.287)	(0.256)	(0.206)	(0.228)	(0.247)	(0.194)	(0.253)	(0.279)
Observations	881	400	404	873	804	777	812	811	808
R-squared	0.510	0.730	0.632	0.648	0.767	0.671	0.515	0.596	0.455

between 12 to 15 years coincides not only with transition to adolescence but also the transition from primary schooling to middle and lower secondary schooling. It is certainly possible that boys and girls are being streamed towards schools of systematically different quality. While this may be captured indirectly in our controls above, e.g. if better schools charge higher fees then perhaps child-specific educational expenditures should have captured part of this difference already, this cannot be assumed necessarily. Therefore in this section we investigate directly the potential for gender-based sorting across schools to explain emergent gender gaps in learning.

Although the Young Lives data used in this paper is based on household visits, with no direct collection of data at the school level, we can uniquely identify the school that each Young Lives child attends (if enrolled) and which other children in our sample attend the same school. Thus, even in the absence of within-school data collection, we can control for school fixed effects in each country.[30] Specifically, we add a full vector of school dummies to Equation (5) but omitting the dummy variable for enrolment;[31] the coefficient of the male dummy variable can thus be interpreted as the remianing gender-based divergence in test scores, conditional on both enrolment and the quality of schools.

Table 9 presents the coefficient on the male dummy variable after the inclusion of school fixed effects. As can be seen, although in no country does school-based sorting succeed in explaining all of the gender-based divergence, the extent to which it can narrow the unexplained portion of the gender-based divergence differs importantly across countries. It is least effective, as a

[30] Conditioning on lagged achievment and basic demographic characteristics of inividuals, this is identical to the procedure for generating school value-added estimates, a procedure for which similar models have been found unbiased in a range of applications recently.

[31] Note that, since some children are not enrolled, we do not need to omit one school in order to identify school fixed effects.

potential channel creating gender gaps, in Ethiopia where the coefficient on the male dummy looks identical to those in previous specifications. In Vietnam, the coefficient on the male dummy variable for math declines by half and is now statistically insignificant and in the cloze test declines by about a fifth, compared to the most extensive specifications presented in Tables 4-6.[32] It is most effective in India where the coefficient on the male dummy variable declines for all of the tests - to about 0.15 SD in mathematics (compared to an unconditional gap of ˜0.35 SD, which had declined to ˜0.25 SD in the most extensive specifications previously), to about 0.2 SD in receptive vocabulary (compared to an unconditional gap of ˜0.4 SD, which had declined to about 0.3 SD in previous decompositions) and to a statistically insignificant negative 0.08 SD in the Cloze test (compared to a positive unconditional gap of 0.23 SD, which had declined to ˜0.12 SD in previous regressions); the coefficients on the male dummy remain statistically significant in both the math and vocabulary tests. In Ethiopia, sorting across schools does not reduce the coefficient at all. [33]

Finally, there is one additional possibility that we have yet to explore. If there is gender-based discrimination in schools, the quality of the same school may yet be different for boys and girls. Such discrimination could well result from e.g. differential encouragement provided to boys and girls, from the 'stereo-typing' effects, or an explicit targeting of resources towards students of one gender. While we cannot observe any of these factors directly, it is possible for us to assess whether schools have a differential effect on the test scores of boys vs. girls. Specifically, we can estimate the following regression:

$$Y_{ia} = \alpha + \beta_1.male_i + \beta_2.male * enrol_{i,a} + \beta_3.Y_{i,a-1} + \beta_4.X_i + \beta_5.EdExp_{ia} + \beta_6.HAZ_{ia} + \theta_s + \epsilon_{ia} \quad (6)$$

$male * enrol$is the interaction of a dumy variable for enrolment with the male coefficient, θ_s is a vector of school fixed effects and all other variables are defined as previously.[34] In this specification, as previously, sorting across schools is captured by the vector of school fixed

[32] We should, however, be cautious in reading too much into the fact that the coefficient is statistically insignificant in some of the tests: the inclusion of a large number of school dummies comes naturally at a cost of statistical power and all estimates are more imprecise than previously.

[33] That the sorting into different quality schools seems to have greater explanatory power for explaining gender gaps in India, than e.g. Ethiopia where school choice is more limited, is consistent with recent work which documents a greater likelihood for boys to be sent to private schools and that private schools have modest positive effects on mathematics and receptive vocabulary (for evidence using Young Lives data, see Singh 2015; see also Muralidharan and Sundararaman 2015).

[34] Note we do not need to include a dummy variable for enrolment given that we already include a dummy variable for every school attended by individuals in our sample(s). Whereas the school dummies capture the productivity of each individual school, the interaction term here captures differences in the average productivity of schooling by gender across all schools; this is also the parameter of interest in our application. In the most flexible specification, each individual school dummy variable could be interacted with the male variable - this is clearly not a feasible alternative in our case given the large number of schools and the small number of individuals per school.

Table 10: Does differential effect of schools account for divergence in learning?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Ethiopia			India			Vietnam	
VARIABLES	Math	Vocabulary	Cloze	Math	Vocabulary	Cloze	Math	Vocabulary	Cloze
Male	0.428***	-0.417	0.296	0.133	0.230	-0.271	-0.0338	-0.0402	-0.261
	(0.160)	(0.332)	(0.205)	(0.172)	(0.161)	(0.202)	(0.106)	(0.139)	(0.160)
Male*Enrol	-0.220	0.695**	-0.299	0.0244	-0.0412	0.296	-0.0288	0.0776	0.105
	(0.167)	(0.342)	(0.225)	(0.196)	(0.179)	(0.231)	(0.119)	(0.156)	(0.173)
Lagged math score	0.386***			0.342***			0.297***		
	(0.0325)			(0.0446)			(0.0400)		
Lagged vocabulary score		0.357***	0.377***		0.531***	0.447***		0.267***	0.213***
		(0.0653)	(0.0629)		(0.0566)	(0.0600)		(0.0362)	(0.0388)
Constant	-0.203*	1.491***	-0.942***	0.346*	1.094***	-1.050***	1.278***	1.674***	-0.808***
	(0.115)	(0.269)	(0.258)	(0.207)	(0.227)	(0.247)	(0.196)	(0.255)	(0.279)
Observations	880	399	403	872	803	776	806	805	802
R-squared	0.511	0.736	0.633	0.648	0.766	0.673	0.518	0.594	0.452

effects; the coefficient on the male variable captures the gender gap among those individuals who are not enrolled at 15 years of age, conditional on covariates; and the coefficient on the interaction variable $male*enrol$ is the gender difference among the enrolled students, conditional on differences arising from sorting across schools of different quality and the various other characteristics. Results from this exercise are presented in Table 10.

The core result is that there is limited evidence of differential productivity of schooling for boys and girls. The male*enrol interaction is rarely significantly different from zero and, thus unsurprisingly, the coefficients on the male dummy variable is mostly unchanged from Table 9.

5 Conclusion

This paper has sought to look at the emergence and evolution of gender gaps in learning over an extensive period of childhood from preschool to early adulthood. Our core contribution has been to use a fresh dataset with comparable test scores available over time and across countries; to document substantial heterogeneity in the direction and magnitude of gender-based gaps in different domains of cognitive achievement in four developing countries. As noted in the introduction, and the analysis presented in previous sections, the major pattern in this paper is about the age trajectory of gender gaps in learning - in no country do we see substantial gender gaps in achievement at early ages, but by the age of 12 years and later we see strong evidence of gender gaps in multiple contexts although with some heterogeneity in whether it favours girls or boys. Between half and two thirds of the cross-sectional gaps at 15 can be explained with recourse to differences in investments, time use and schooling. However, a substantial portion of the gender gap remains still to be explained.

References

Andrabi, T., Das, J., Khwaja, A. I., & 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., & Walters, C. R. (2013). Explaining Charter School Effectiveness. *American Economic Journal: Applied Economics*, 5(4), 1–27.

Bharadwaj, P., De Giorgi, G., Hansen, D., & Neilson, C. (2012). The gender gap in mathematics: Evidence from low-and middle-income countries. *NBER Working Paper*, (w18464).

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

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

Deming, D. J. (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review*, 104(5), 406–11.

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

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

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. *Handbook of labor economics*, 4, 1–102.

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

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

Grant, M. J. & Behrman, J. R. (2010). Gender gaps in educational attainment in less developed countries. *Population and Development Review*, (pp. 71–89).

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human resources*, 45(4), 915–943.

Jensen, R. (2012). Do labor market opportunities affect young women's work and family decisions? experimental evidence from india. *The Quarterly Journal of Economics*, 127(2), 753–792.

Kane, T. J., McCaffrey, D. F., Miller, T., & 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*.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *The Journal of Economic Perspectives*, (pp. 167–181).

Munshi, K. & Rosenzweig, M. R. (2006). Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy. *The American Economic Review*, 96(4), 1225–1252.

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

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

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