Diagnosing the Learning Crisis: What Can Value-Added Analysis Contribute?

Abstract

Advocates of teacher value-added modelling (VAM) argue that this technique can provide evidence on teacher effectiveness to inform teacher policies and broader education system reforms. Critics contend that value-added is a poor proxy for teacher quality and as such is of questionable utility, especially where teacher accountability is concerned. In low- and middle-income countries, and especially sub-Saharan Africa, where the challenge of the 'learning crisis' is most severe, a lack of longitudinal data has precluded extensive debate on the matter.

In this paper we explore the potential of value-added analysis for diagnostic purposes in the context of Ethiopia. We make use of data from the Young Lives longitudinal study – specifically two rounds of school surveys conducted in Ethiopia between 2012 and 2017 when pupils were in grades 4 to 8. Learning levels in the Young Lives sites in Ethiopia are very considerably below curricular expectations. Like many countries in sub-Saharan Africa, Ethiopia faces a significant challenge in terms of a 'learning crisis' and in terms of the attendant need to develop policies to improve educational effectiveness within the confines of very limited resources. We discuss the background to VAM models and their use, including in relation to the context of Ethiopia.

The paper shows that learning progress in primary schools varies widely between classrooms, and between pupils within the same classroom. Some schools and teachers are more successful in raising overall attainment by 'raising the floor' of learning and narrowing the dispersion. Others are more successful by 'raising the roof'. Less effective teachers appear to be particularly ineffective for pupils with higher scores at the start of the year. In contrast, the most effective teachers showed high levels of 'value-added' for pupils at all levels of prior performance. Diagnostic analysis of teacher value-added has potential, we argue, to aid understanding of contributors to low levels of learning such as: (i) over-ambitious curricula; (ii) absence of 'teaching at the right level'; (iii) within class heterogeneity and pupil grouping strategies; and (iv) teaching and learning strategies – such as 'differentiation' or 'mastery'.

Key words: Teacher value added; raising learning achievement; Ethiopia; Learning gains.

1. Introduction

The crisis of low levels of learning in Ethiopia, as in many sub-Saharan African contexts, affects a majority of pupils (see World Bank 2016). Clearly the causes of low levels of learning outcomes are many and include scarcity of resources in education and disadvantage at the level of pupils' home backgrounds; including disadvantage linked to low average levels of education among parents and the wider community. Indeed, in less economically developed regions of Ethiopia, many pupils entering into primary school are 'first generation learners' (see Iyer et al 2020). Despite these powerful influences on learning, the literature suggests that differences in 'effectiveness' between schools and teachers nonetheless play an important part in determining pupils' learning progress, while much of the evidence continues to derive from studies in high income countries (see Scheerens, 2000). At the same time, evidence on exactly which features of schools or teachers and their teaching are more effective in particular low and middle income contexts is both scarce and somewhat inconclusive; in part because 'what works' in terms of particular pedagogical strategies, for example, has rather limited generalisability (see Glewwe et al., 2020). In this paper, we do not address the question of what, in pedagogical terms, lies behind effective teaching, but instead focus on what differences in teacher effectiveness (or 'value-added') as a whole can tell us at a diagnostic level about the organisation (and potential remediation) of an education system, in our case, Ethiopia's system.

In Ethiopia, teachers are expected to deliver a grade-specific curriculum, set out in textbooks and teachers' guides (see for example USAID, 2017). The curriculum is designed to be cumulative in grades, with higher-order skills building on their respective foundations. However, due to slow learning progress in early primary grades, language-based transitions in upper primary grades and automatic promotion policies; alongside socio-economic disadvantage and limited support at home, many children reach grade seven or eight substantially lacking in the prerequisite skills; and may be several grades behind in terms of their understanding of the 'expected' curriculum (see Rossiter et al., 2018). Without special attention, we may expect the learning progress of these pupils to be limited by the mismatch between curriculum content and their developing skills base (i.e. excessive distance from the zone of proximal development in Vygotsky's terms). Skilled teachers will, of course, be able to make adaptations in the classroom, to address and mitigate the gap between pupils' skills and curriculum expectations. This may be more demanding for less well pedagogically trained or less experienced teachers and where classes are large and poorly resourced; each of which are pervasive issues in Ethiopia (see Barnes et al., 2018). Indeed, adaptation may seem impossible or futile in the most demanding of contexts – very large and heterogeneous classes with inadequate resources and where knowledge of pupils' actual learning levels is weak in the absence of adequate assessment.

Where possible, however, remediation strategies might include providing extra support for low-achievers early-on to enable them to 'catch up' and to access the expected grade-specific content, in addition to 'differentiation' of learning tasks and expectations by teachers within regular classes to cater to a variety of learning needs. Some specific interventions focused on intensive supportive remediation have shown very large effect-sizes in low- and middle-income contexts (see Snilstveit et al 2016 for a review), including where curricula have been described as 'over-ambitious' (Pritchett and Beatty 2012), most notably in India (see for example Banerjee et al., 2016). These large effects may attest not only to effective intervention design and implementation but also to the sheer extent of the 'learning gap' between pupils who are 'on target' and those who are 'behind' and to the failure to address it within the existing mainstream education system, rendering this an 'open goal' for intervention.

The gap between curricular expectations and actual attainment levels may be exacerbated when teacher training and incentives are strongly aligned with 'delivering the prescribed curriculum' and when teachers may, accordingly, perceive adaptation as risking the progress of high achievers whose skills do ordinarily keep pace with expectations. Where the gap is large and teachers effectively 'teach to the top', the effect may be to further increase dispersion of pupils' learning outcomes, raising both inequality and inequity, while preparing a small proportion of, typically advantaged, pupils for progression to the next stage of education. This last aim is arguably consonant with the historical structure of many education systems globally, which were not designed to equip children with skills *en masse* but to select and prepare an elite. In Ethiopia, under the imperialist system until as recently as 1974, more than 90% illiteracy prevailed in a system which served only a small minority (Gupta, 1994).

However, while reducing inequality (dispersion) in learning levels is desirable, it is especially desirable when this reduction comes from 'raising the floor' in terms of attainment levels (see Crouch and Rolleston, 2017; Crouch et al., 2020), rather than at the expense of pupils who are on target to succeed. Enabling teachers to 'add value' and raise learning outcomes for low and middle performers is especially crucial in contexts such as sub-Saharan Africa, where few pupils' learning outcomes meet either national curricular expectations or international benchmarks as provided by studies such as PISA and TIMSS (see Pritchett and Viarengo, 2021).

Appropriate data for examining teacher value-added and especially differential value-added; that is the extent to which 'overall effectiveness' is derived from effectiveness for particular groups of pupils, is limited in sub-Saharan Africa. One source of data which does permit this kind of analysis is the Young Lives longitudinal study (discussed in Section 3), which includes data from households and schools collected over a period of 19 years to date in four countries, including Ethiopia. In this paper we employ data from the Young Lives school surveys to examine the prevalence of low learning outcomes and the 'learning gap' between curricular expectations and actual measured learning outcomes. We then proceed to present the results of value-added models for particular groups of pupils in terms of the distribution of outcomes in a mathematics test conducted among pupils in Grades 7 and 8 in the Young Lives Ethiopia School Survey.

We address the following questions:

(i) How do pupils' actual learning outcomes in Ethiopian primary schools in the Young Lives sample compare to curricular expectations?
(ii) To what extent does differential teacher effectiveness play a key role in explaining the progress of groups of pupils with different starting points?
(iii) To what extent do teachers of different levels of 'overall effectiveness' tend to achieve this effectiveness by raising learning outcomes of particular groups of pupils?

The paper is structured as follows. Section 2 reviews the literature on 'teacher value-added'. Section 3 outlines the data and methods employed, specifically the value-added approach and the Young Lives Ethiopia data. Section 4 provides descriptive analysis of learning outcomes data for Ethiopia while Section 5 presents the evidence on teacher value-added and differential value-added. Section 6 discusses the results and their implications and concludes.

2. Teacher Effectiveness and 'Value-Added' Measures

The use of value-added modelling (VAM) has become a norm in some, mostly developed, contexts for measuring teachers' effects on pupil learning. Instead of focusing on linking specific observable characteristics of teachers to pupil outcomes as in a basic 'production function' approach, VAM uses changes (progress) in pupils' test scores to estimate teacher 'quality' or 'effectiveness' in terms of what may be considered a 'black-box' effect. Accordingly, an 'effective teacher' is one who consistently produces above average gains in pupils' test scores. The concept of teacher 'value-added' can be traced back to the 1970s, when the VAM approach was first used to infer teacher quality (Armor et al., 1976; Hanushek, 1971; Murnane, 1975). It has recently received renewed interest and focus from researchers and policy makers, especially in the United States, partly due to increased availability of longitudinal data from school systems. VAM has gained increasing attention in public policy debates such as those concerning teacher and school evaluation and improvement (Hanushek & Rivkin, 2012; Isenberg & Hock, 2010).

VAM may be employed for a number of distinct but related purposes, including school and teacher accountability, providing information for school choice and for education system diagnostics. These purposes may require different approaches in statistical terms – especially concerning adjustment for covariates such as pupils' backgrounds and school resources. While adjusting for pupils' backgrounds, for example, may be 'fairer' for the purposes of teacher accountability, this is controversial when providing information for school choice because it may lead to at least a perception that lower expectations are being set for more disadvantaged schools and pupils. Interpretation of VAM results may also vary according to the purpose, for example, with regard to whether specific identification of causal effects is required. Questions around whether VAMs are able to provide unbiased causal estimates of teacher effects on pupil outcomes depends partly on the choice of model and covariates included in the model as well as the extent of any issues of measurement error and the efficacy of strategies employed to address this. Todd and Wolpin (2003) provide a seminal discussion of a number of these issues.

Value-added models can be specified differently depending on the intentions and assumptions made by researchers. McCaffrey et al. (2003) separate VA models into univariate and multivariate outcome models. While univariate models focus on one outcome (for example test-score) per model, multivariate models allow multiple years of outcomes to be modelled jointly. Even though the multivariate approach may be more flexible and efficient, Everson (2017) shows that the majority of value-added research conducted during 2007-2015 primarily used univariate models. In the univariate approach, choices may be made between using either 'gain-score' or covariate-adjusted models in which a prior (lagged) test-score is used as a covariate. Gains models may be considered the simplest form of VA models (Everson 2017: 40). Gains or the differences between pre- and post-test scores are used as an outcome variable. The gains can then be linked to teachers or schools. However, models using gains may be limited when the two tests are on different scales. Hence, covariate-adjusted models are typically employed in these cases. Everson (2017) shows that the majority of research studies on teacher VA for accountability employ this method in preference to the gains model. This may also be because the specification fits with the education production function framework which describes pupils' current achievement to be a function of different factors (Hanushek & Rivkin, 2010) and this model allows for more flexible modelling of 'skill retention', allowing for atrophy of skills or forgetting of knowledge over time. Consequently, most studies using VA include at least some form of covariate modelling, including at least a prior test score.

While pupils' prior test scores are included in some way by definition in VAM, it is arguable which other variables should be included in the estimation of value-added. Technically, school and teacher value-added estimates can be computed using only current and prior test scores (no other covariates), which may be termed 'unconditional' VA estimates. However, by adding variables such as indicators of pupils' backgrounds, the estimates become conditional or adjusted, based on the variables included. Theoretically, this may help capture the differences in pupil selection into schools and classrooms (Everson, 2017). In other words, as pupils and teachers are not randomly assigned to schools, estimated teacher VA could be the result of both the true teacher effect on achievement and/or other unobserved factors that affect achievement; such as 'sorting' of pupils with different abilities or motivations into schools and classes via mechanisms such as school choice and selection. Without properly controlling for sorting, high or low teacher VA could be the result of unobserved differences linked to the allocation of pupils to particular schools and classes rather than an estimate of 'true' teacher effects (Chetty, Friedman, & Rockoff, 2014). In which case, the VA estimates obtained would be biased. To counter this, McCaffrey et al. (2003) argue that explicitly controlling for variables that predict sorting in the model may reduce biases from omitted variables (McCaffrey et al., 2003). Specifically, where the goal of VAM is to isolate teacher or school effect on achievement (e.g. especially for teacher accountability), the variables that predict achievement but are not related to teachers should be controlled for. Covariates used by researchers include pupils' demographics (Backes et al., 2018; Chetty et al., 2014; Rothstein, 2010; Stacy, Guarino, & Wooldridge, 2018), proxies for socioeconomic status (Stacy et al., 2018), and occasionally peer effects (Harris, Ingle, & Rutledge, 2014). The specific variables included typically include gender, age, grade repetition, ethnicity and socio-economic indicators.

Empirically, research in support of using pupil background information includes Dearden, Miranda, and Rabe-Hesketh (2011) who found that mother's education is highly related to pupils' scores and that adding mother's education as a covariate significantly changes the estimated value-added. Hence, they concluded that the value-added models that do not account for this variable may be considered biased. Nonetheless, others argue that covariates may not be necessary in order to obtain valid VA estimates. In the Tennessee Value-Added Assessment System (TVAAS), one of the largest teacher assessment programmes, background variables are not explicitly controlled for. Ballou, Sanders, and Wright (2004) show that adding socioeconomic and demographic variables does not significantly change the VA estimates in the case of TVAAS. Similarly, Chetty, Friedman, and Rockoff (2014) found that prior test scores of pupils are the key covariates that makes the value-added estimates unbiased. These prior test scores may be expected to have absorbed effects of a host of child, family background and prior school and teacher level influences on attainment, especially where the prior test score is relatively recent. Chetty, Friedman, and Rockoff (2014), however, do show that adding pupils and parents' characteristics can also improve the model but that failure to include them does not significantly bias the results. So far, there is not general agreement on which controls should be included in the model and this rather depends on the purpose. This may lead to researchers instead relying on data availability rather than theoretical foundations in the inclusion or exclusion of certain variables (Everson, 2017). Therefore, results based on VA models may be sensitive to how the models are specified. Additionally, they may show significantly different results when the context of the study is changed.

The purposes for which value-added estimates may be used are, potentially, many and various, while perhaps teacher accountability is the most common and best known as well as being most controversial. Issues concerning the reliability and validity of estimates are particularly acute when used for individual accountability. Kupermintz et al (2001) discuss the difficulties in relation to the TVAAS while Darling-Hammond (2015) proposes a cautious approach to teacher evaluation, employing value-added estimates alongside a range of other evidence. Value-added estimates can also be used without linking to individual teachers in broader school effectiveness research and policy analysis. Jerald (2009) discusses the application of value-added data by schools and districts to school improvement more generally, including improving targeting, efficiency and equity of policies and interventions including teacher professional development. Giffin et al. (2009) illustrate the potential uses of value-added data for classroom diagnostics, for example for the analysis of which groups of pupils make more or less progress under particular conditions. Making use of Young Lives data, Rolleston and Moore (2018) employ value-added analysis to identify differences in effectiveness of particular school-types in Andhra Pradesh and Telangana, India as well as examining who benefits from attending more effective schools. They show that more advantaged pupils typically gain access to more effective (often private) schools, identifying a particular source of inequity. Rolleston et al. (2013) undertake similar analysis for Vietnam, connecting value-added estimates to characteristics of schools and teachers. While it is not straightforward to demonstrate causal linkages, descriptive data provide important evidence for potential diagnostic analyses. Their study shows, for example, that classes where classrooms, teachers and pupils lacked materials were associated with weaker value-added, while classes taught by teachers with positive attitudes towards pupils' learning progress and teachers who were more often evaluated were associated with higher value-added. While there can be no simple read-off from value-added models to policy reform or intervention, the evidence may be used to develop hypotheses and explore potential pathways.

The majority of research on VAM has been conducted in the US, where VAM is used extensively in teacher evaluation. Even though general results show some similarities, the magnitude of average teacher effects varies significantly from study to study. Hanushek and Rivkin (2012) show that research consistently finds large variations in teacher quality using VAM, including within-schools. This indicates that teachers differ in their ability to improve pupil outcomes and this is evident even with teachers from the same school. They summarise the variation in teacher effectiveness in selected studies using standard deviation of teacher 'fixed effect' estimates, expressed in terms of pupil outcomes. These numbers (averaging 0.13 in reading and 0.17 in mathematics based on a normalized test score with mean 0 and standard deviation 1) may be interpreted as measures of the potential impact of increasing teacher effectiveness. For example, an increase of one standard deviation in teacher VA may be expected to result in an increase in pupil learning of 0.13 standard deviations in reading based on this average estimate. Chetty et al. (2014) go further, to argue that, based on the improvement in lifetime earnings which is linked to better educational outcomes, an increase in teacher 'quality' in terms of value-added of one standard deviation could translate, in their US example, into a lifetime increase in earnings of as much as $39,000.

This illustrates the significance of teachers in contributing to pupils' academic outcomes and raises important questions about the importance of policies regarding how teachers are allocated to schools and pupils and how teacher effectiveness may be improved overall or how poor effectiveness may be addressed. However, some researchers report notably higher teacher effects than others so there is not a clearly consistent picture in this regard. For instance, the average teacher VA is consistently higher in reading comparing to mathematics. Value-added reported by published research during 2004-2010 comprised by Hanushek and Rivkin (2010) shows the range of mathematics VA to be from 0.08-0.26 standard deviations while it is 0.11-0.36 in reading.

To assess the value of VA estimates, precision and stability are two of the criteria that may be used. Stacy et al. (2018, p. 51) define precision as "estimates of the variance of estimation error, or squared standard errors, of value-added measures". Here, the estimates are more precise when the standard errors are small. With smaller standard errors, the confidence intervals in which the true estimates locate are smaller as well. Precision is an important feature for VA estimates to have, especially when it is to be used for teacher evaluation (McCaffrey et al., 2003). As for stability, Stacy et al. (2018:51) define this as "correlation from year to year in a teacher's value-added measure". A stable estimate is an estimate that does not vary substantially from year to year. In addition to a time factor, Everson (2017:52) extends this definition to also include stability over testing instruments or models used for estimation. Both precision and stability could be used as an evaluative tool for VA models. Yet, the interpretation of precision and stability of the estimates should be used with care. This is because unstable estimates may not necessarily invalidate VA results as the change over time may be from external factors or real changes in teacher 'ability' or effectiveness (Everson, 2017; Stacy et al., 2018).

For precision, researchers generally agree that as sample size increases, the estimates become more precise (Everson, 2017; Gulosino, 2018; McCaffrey et al., 2003; Stacy et al., 2018). Hence, VA estimates may be less precise for teachers or classrooms with small number of pupils. Other research investigating the issue focusses on whether precision changes with different groups of pupils. Stacy et al. (2018) explored whether the estimates are similarly precise for teachers serving differentially performing pupils. They found that for all grades (grade four and six) and subjects (Mathematics and English) investigated in the study, the standard errors of the VA estimates are significantly higher for teachers who teach pupils with performance at the bottom 25% of the sample. Even when the number of observations is set to be equal for both groups of pupils, the gap in precision remains for the high- and low-performing pupils. This implies that teachers who are assigned lower-performing pupils have less precise or more variable VA estimates. Similarly, Herrmann, Walsh, and Isenberg (2016) found that precision is lower for teachers who serve pupils of more disadvantaged background (having lower prior scores and eligible for free lunch). From the findings, it seems that precision of VA estimates can be affected by pupil composition. Additionally, dimensions of classroom composition are also found to affect the value of the estimates themselves. Horoi and Ost (2015) found that classes with more 'disruptive' pupils return lower teacher VA estimates. Therefore, both the differential estimates and precision of estimates observed between studies may partially be driven by classroom composition.

VA estimates have been examined in terms of their stability over time, model, and instrument. In terms of stability over time, most research agrees there is some degree of instability of the estimates across time. Yet, the magnitudes differ from studies to studies. Ferrão (2012) examines achievement data from one region of Portugal. Even though the VA estimates are relatively stable over two years, with 65% of teachers located in the same quartile rank, only 12% of the teachers remain in the same rank over three years. Other researchers have investigated stability over longer period of time. McCaffrey et al. (2009) estimated the stability over five years using data from different districts in the US. They found low to moderate correlations of year-to-year VA, with middle schools having relatively higher correlation than primary schools. Similarly, Goldhaber and Hansen (2013) argue that VA based on smaller number of years may not be stable as the data is noisy due to many factors including measurement error, non-persistence fluctuations, or dynamic changes in performance. Based on 10 years of data, they concluded that the part of VA that is fixed due to teacher quality is relatively small (29%) in relation to other variations. Similar conclusions have been drawn from qualitative work. Close and Amrein-Beardsley (2018) present accounts of a teacher who receive significantly different VA ratings in two years despite having similar mix of pupils and teaching methods. Lack of stability over time may indicate that teacher value-added estimates may include other (unobserved) factors in addition to 'true' teacher quality. These unobserved factors, if not controlled for, may cause results from VA models of different years to differ.

In addition to stability over time, some studies examine stability when employing different modelling approaches. Kurtz (2018) shows that VA estimates are unstable across model specifications. Specifically, he compared between value-added model which estimates teacher effectiveness using mean scores and pupil growth percentile model which uses median. Across the two models, the presence of large number of pupils who have either very high or low growth is found to be linked to significant differences in estimates. This implies that if a teacher is assigned to pupils that deviate from the mean in terms of growth, he/she may get very different VA results. This appears to be consistent with the instability of VA across specifications. Backes et al. (2018) took a different approach to evaluating stability. They utilised a natural experiment of assessment regime changes in several states in the US to investigate whether the stability of VA changes during this period or not. In contrast to Stacy et al.'s (2018) findings, the VA estimates are not significantly different in advantaged and disadvantaged classrooms (having high or low prior scores). However, the results have more volatility in the case of disadvantaged classrooms. They also found VA to be relatively stable during assessment changes in mathematics, but less so in reading.

Despite some generally consistent findings, estimates from different VA research vary in terms of magnitude, precision, and stability. Specifically, pupil composition, grade level, and subject seem to partially explain the differences in the estimated VA. It is also important to note that the majority of VA research has been conducted in the US. Hence, using VA in a different context may yield different findings as well, especially in developing countries where diversity in schools and classrooms may be more apparent. The next section explores the ways in which VA may be used in the context of Ethiopia, using Young Lives dataset.

3. Data and Methodology

Young Lives is a longitudinal study of childhood poverty in Ethiopia, India, Peru, and Vietnam. It has followed a total of 12,000 children over the course of 19 years. In all four countries, a sentinel-site sampling design[1] is employed, comprising 20 purposively selected sites in each country. Full details are available in Boyden and James (2014). In 2010, a school component was introduced to explore Young Lives children's experiences of schooling and education in depth. Three school surveys have been conducted in Ethiopia: at lower primary level in 2010 and 2012-13, and at upper primary level in 2016-17. School surveys sample Young Lives index children and their peers in classrooms across the original 20 sentinel sites in Amhara, Tigray, SNNP, Oromia and Addis Ababa, and in an additional 10 sites in Somali and Afar regional states. The final sample is not nationally representative but the sites are selected to broadly represent national diversity, with a pro-poor bias (excluding the most advantaged areas of the country). At site-level the samples of children in the Young Lives longitudinal household study are randomly selected from within a birth cohort. School surveys, however, include all the children attending school in the selected school grades in schools within the study sites; in other words they include the peers of the Young Lives index (birth cohort) children.

In this paper we use data from the Ethiopia school surveys conducted in 2012-13 (Grades 4 and 5) and in 2016-17 (Grades 7 and 8) to answer research question (i). By including both rounds we can review learning levels against curricular expectations across the primary school cycle. For research questions (ii) and (iii) we focus on the 2016-17 survey data collected at the beginning and the end of that school year. All survey modules were completed on paper by school headteachers, classroom teachers and pupils, with support from trained enumerators at each site. Full details regarding the design, implementation and results of these surveys are available in Aurino et al (2014) and Rossiter et al (2017).

Each survey included linked mathematics tests administered at the beginning and end of the relevant school year. Assessment items were developed in collaboration with curriculum experts at Ethiopia's Federal Ministry of Education, with items linked to documented curricular expectations for the relevant grades. Tests contained a number of common items to allow concurrent calibration of latent-trait scores using item-response modelling so that test scores from both test waves are reported on a common scale. The mean test score was defined at 500 in the first wave (beginning of the year) test and the standard deviation at 100. Azubuike et al (2017) explain the design and piloting of test instruments, along with assessments of reliability and validity for their application in school surveys.

The value-added approach we follow to produce the estimates in Section 5 is a simple extension to the basic education production function which uses an end of year test score as the outcome. It introduces a measure of prior attainment (i.e. the beginning of the school year test result), to account for the contribution of all relevant prior educational inputs whose effects are reflected in a pupil's attainment at the time of the test. A full discussion of the value-added framework is provided in Todd and Wolpin (2003).

Equation (1) below describes the general framework in simple terms. T represents a test score of pupil i at time t. X is a vector of child characteristics and η an individual error term. The framework may be extended to include school/teacher characteristics or school/teacher effects, modelled as fixed or random effects.

[1] An approach to non-representative site sampling often used in health surveillance surveys. See Boyden and James (2014:15) for details.

$$T_{i,t} = \alpha T_{i,t-1} + \gamma X_i + \eta_i \qquad (1)$$

In this paper we use a three-level multi-level model (MLM) in order to model the effects of schools and teachers in a hierarchically structured dataset – pupils nested within classes (teachers) nested within schools as denoted in (2) below. In this equation, k denotes the pupil-level, j the class-level and i the school level; u_i denotes a school-level (random) effect, u_{ij} a class-level (random) effect and η_{ijk} an individual (pupil-level) error-term.

$$T_{ijk,t} = \alpha T_{ijk,t-1} + \gamma X_{ijk} + u_i + u_{ij} + \eta_{ijk} \qquad (2)$$

The MLM approach allows us to understand how much variation in learning outcomes arises at school versus class (teacher) and pupil levels and how this is affected by adjustment for covariates as well as taking account of this structure in the estimation. We focus on the class (teacher) level. Table 0 reports the intra-cluster correlation co-efficients for the total sample model in order to illustrate the extent of clustering in the data. First of all, the ICCC from an 'empty' model with no covariates (variance partitioning model) is reported; secondly for a model with prior attainment in mathematics and English included as the only covariates and thirdly with the full set of pupil explanatory variables and controls as described in Section 5. The results show a relatively high degree of clustering at both class and school levels in terms of end of year mathematics outcomes. Clustering is reduced substantially when taking account of prior test scores and even more so with the full set of explanatory variables and controls. This clustering suggests that MLM is appropriate but also indicates that there is notable within school homogeneity and between school heterogeneity in terms of attainment. The aim of the MLM exercises is not specifically to attribute causal relationships but to describe the sources of variation in mathematics progress for several groups of pupils, with particular attention to the class/teacher level.

After estimating a model for all pupils, we estimate separate models for five quintiles of pupils based on their attainment in mathematics at the beginning of the school year, using a test score at the end of the school year as the outcome. In our analysis we focus on class (teacher) level effects. Note that teachers in a particular school may teach more than one class and typically teach two classes or more (often one class in Grade 7 and one in Grade 8). We include each class separately, however, on the basis that teachers may be more effective with one class than another, linked to issues including class composition as well as pedagogical strategy. Accordingly, our analysis is of teacher-class combinations. The survey sample includes 9366 pupils (in five quintiles of 1862/1863 pupils) in 250 classes within 56 schools. Descriptive statistics on each of the quintile samples are available in Appendix tables A1-A5.

Table 0: Intra-Cluster Correlation Co-efficients

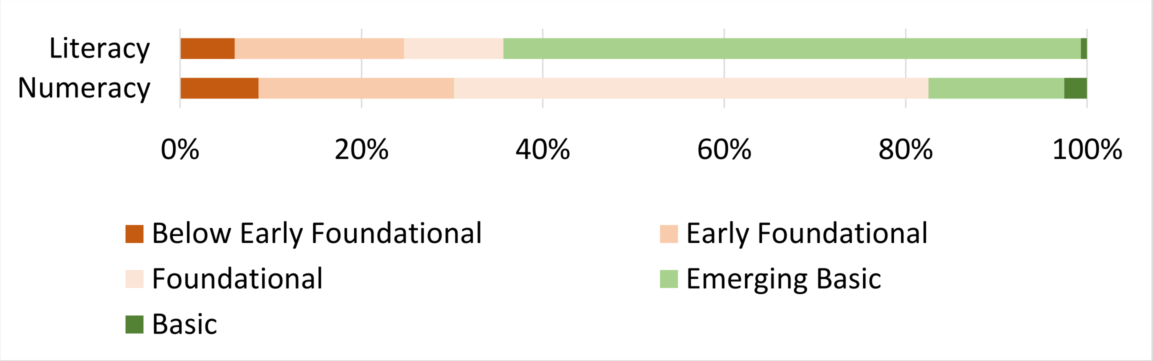
	Empty Model	*Prior Attainment Only*	*Full Model*
School	0.30 (0.05)	0.06 (0.02)	0.01 (0.01)
Class	0.38 (0.04)	0.11 (0.02)	0.06 (0.01)

Standard errors in parentheses.

4. Learning Outcomes and Curriculum Expectations in Ethiopia

Figure 1 below employs data from the Young Lives primary school survey in Ethiopia conducted in 2012/13 in Grades 4 and 5 with pupils typically aged 11. Young Lives conducted a benchmarking exercise to map national curricular competencies to test data from the survey and created five indicative skill groupings as shown below. The curriculum target or expectation in terms of skills for these grades is benchmarked here as 'basic' or at the very minimum 'emerging basic' skills. A large proportion of pupils, especially in numeracy, have clearly not reached these levels. Only a tiny minority had unambiguously mastered the intended skills.

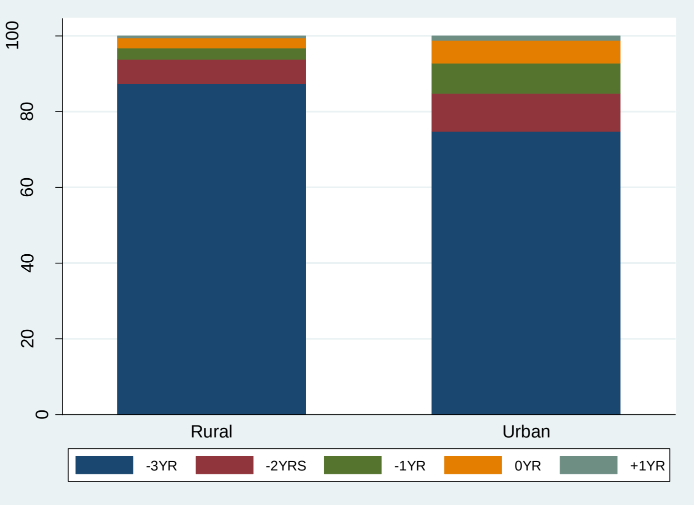
Figure 1: Pupils Achieving Basic Skills at age 11 in Ethiopia (2013-14)



Source: Authors' computations using Young Lives data (2012-13)

The Young Lives Upper Primary School Survey, conducted in 2016-17, followed pupils in the same age cohort (typically aged 15 by this time) and included beginning- and end-of-year mathematics assessment data for 9,434 pupils in Grades 7 and 8 in 271 classes across 63 schools. Figure 2 below shows a similar picture to Figure 1 regarding learning levels four years later. This time we report the number of school years behind curricular expectations which corresponds to pupils' actual assessed learning levels in Young Lives' tests. The vast majority of pupils' learning is found to be at least 3 years behind expectations, both in urban and rural areas of the country.

Figure 2: Pupils' Learning Outcomes in Maths at age 15 in Ethiopia, Compared with Curricular Expectations (2016-17)

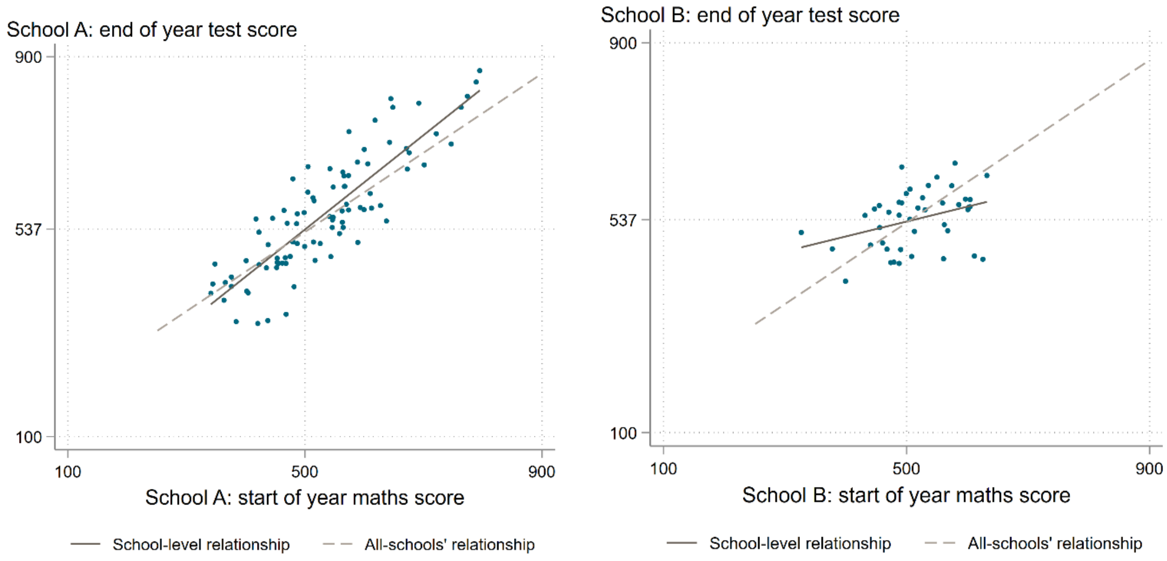


Source: Authors' computations using Young Lives data (2016-17)

5. Learning Gains and Teacher Value-Added

Figure 3 depicts two purposively selected schools with somewhat different groups of pupils and somewhat different patterns of learning progress. These are actual schools in the Young Lives 2016-17 survey (Grades 7 and 8). School A demonstrates substantial progress for pupils with above average prior achievement and limited progress for pupils with below average prior achievement. By contrast, School B shows a more homogenous group of pupils (in terms of prior achievement), with more progress among those with below average prior achievement – perhaps at the cost of high-early-achievers' progress. School A is in fact a school with greater than average 'value-added' (effectiveness) while School B is closer to average, while it is clear that the two schools are effective by improving learning among rather different groups of pupils – school A 'raises the roof' while school B adds value more evenly across the range of prior learning levels.

Figure 3: Learning Gains in Two Schools: Scatter Plot of Beginning and End of Year Test Scores[2]



School A

School B

Source: Authors' computation from Young Lives data

Using value-added estimates, which are generated separately for pupils by quintile of performance on the beginning of the year test, we are able to examine the question of 'for whom' teachers are effective. In our approach, pupils are divided into quintiles of prior performance, Q1 to Q5, where Q1 is the lowest performing 20% of pupils from the beginning of the year. As classrooms are not groups of homogenous average learners, teacher effectiveness may vary according to many pupils' characteristics, including ability and prior learning. As seen above, single estimates of teacher value-added based on a particular (whole) class may mask this heterogeneity in effectiveness. For this exercise – and in contrast with the curriculum benchmarking reported in Figures 1 and 2 – we use achievement scores estimated using item-response analysis. This provides an estimate of each pupil's achievement on a fine-grained scale with mean 500 and standard deviation 100.

[2] The individual school-level relationship in each case is denoted by the dashed line and the whole sample-level relationship is denoted by the solid line.

Table 1 reports the results of the regressions for each of five quintile group of pupils defined by their mathematics performance at the beginning of the school year, reporting the pupil-level co-efficients. It also includes the results of a model employing data from the total sample. The models include covariates for attainment at the beginning of the school year in mathematics and in English, for age, gender, parental literacy, household wealth, nutrition and attendance at pre-school. We also include controls for region, school ownership type and grade attended (not reported).

Descriptive statistics for the variables included are reported in tables A1-A5 in the Appendix. In summary, older pupils (within the same grade), with the exception of pupils in quintile 2, perform significantly less well at the end of the school year (T2) in mathematics other things equal; including taking account of their performance at the start or the year; indicating that they make less progress. Girls when compared to boys (except in quintile 2) perform less well at the end of the year, controlling for performance at the beginning of the year, as indicated by the positive and significant co-efficient on male gender. Pupils who had attended pre-school in the total sample and in quintiles 1, 4 and 5 made less progress. No significant effects of household wealth or parents' literacy are found in these value-added models. The effect of nutrition is positive and significant in the model for quintile 2 and in the total sample.

The models estimate effects on learning progress over a single academic year, so these results should not be taken to suggest that factors such as home backgrounds are not strong influences. The lagged (prior) test-score variable may be expected to have absorbed many of the longer-term background influences and the regression results may be considered to reflect the influences which are particularly important only during the academic year in question (Grade 7 or 8). Attendance at pre-school may be expected, for example, to increase a child's readiness to enter Grade 1 and may be expected to be associated with higher performance in early grades, but as our models control for prior performance we would only expect an effect of pre-schooling in Grade 7 or 8 if the benefits of pre-schooling exerted a continuing and contemporaneous influence in these grades, rather than a historical one. A similar explanation likely explains the lack of significant effects of parental literacy and household wealth. One possible explanation of the apparent negative influence on progress of pre-school attendance of some groups of pupils is that, although pupils who have attended pre-school have higher levels of performance in absolute terms, there is a degree of 'catch-up' especially on more basic test-items by pupils who did not attend pre-school and who start from a lower baseline.

Table 1: Value-Added Model Results by Quintile of Baseline Test Score in Maths: Outcome Maths score T2 [3]

	Total Sample	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Maths score T1	0.62	0.33	0.44	0.63	0.63	0.64
	(63.16)***	(7.31)***	(4.13)***	(6.21)***	(7.39)***	(20.45)***
English score T1	0.26	0.21	0.26	0.25	0.26	0.26
	(23.38)***	(8.03)***	(9.92)***	(10.21)***	(10.89)***	(10.79)***
Age	-1.92	-2.39	-0.72	-2.38	-2.29	-3.53
	(-3.62)***	(-2.52)**	(-0.68)	(-1.91)*	(-1.83)*	(-2.54)**
Male	6.85	5.56	7.53	2.016	8.09	6.98
	(4.94)**	(2.05)**	(2.58)***	(0.65)	(2.50)**	(2.11)**
Mother reads	-1.50	-2.67	1.01	-3.38	-3.55	0.77
	(-0.92)	(-0.89)	(0.30)	(-0.94)	(-0.91)	(0.18)
Father reads	2.75	5.07	4.48	4.18	-1.97	-3.29
	(1.58)	(1.62)	(1.26)	(1.10)	(-0.47)	(-0.74)
Wealth Index[4]	0.59	-0.61	0.65	1.75	1.74	-0.27
	(1.11)	(-0.68)	(0.64)	(1.53)	(1.40)	(-0.18)
Three meals[5]	3.81	2.45	6.07	5.23	5.02	3.94
	(2.04)**	(0.72)	(1.65)*	(1.28)	(1.11)	(0.81)
Attended pre-school	-5.33	-5.48	-2.48	-1.31	-6.33	-9.01
	(-3.33)***	(-1.90)*	(-0.78)	(-0.37)	(-1.68)*	(-2.17)**
Constant	103.71	271.10	168.26	115.87	106.91	124.02
	(9.93)***	(10.53)***	(3.38)***	(2.16)**	(2.10)**	(4.11)***
Observations	8,270	1,587	1,620	1,651	1,688	1,724
Number of schools	56	53	54	55	56	54
Number of classes	254	225	238	250	244	217

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2 reports the range of school and class level effects (random effects parameter estimates) from the five multi-level regression models. It shows that within each quintile group, class (and teacher) level (within school) effects tend to be larger than school-level effects (within regions and school-types). After controlling for prior attainment and a range of pupil and contextual factors, there remained ‘residual’ class-level effects on progress in mathematics which are most notable in the higher-scoring quintile groups. For example, in the highest-scoring quintile group (5), a class (teacher) which is one standard deviation more effective than average in the overall distribution is associated with an improvement of 19.54 test score points (approximately 0.2 standard deviations) at the pupil level, indicating a relatively high level of heterogeneity in class or teacher effects.

[3] Control variables (not shown) are included in the models for region, school type and grade.

[4] Wealth index created using principal components analysis of household portable assets

[5] Child reports eating three meals per day as opposed to two or fewer.

Table 2: Random Effects Parameters: School and Class Effects

Quintile	1	2	3	4	5
School-level effect SD[6]	7.91	8.47	0.00	8.66	7.69
Standard error	(2.48)	(2.65)	(0.00)	(3.93)	(4.43)
Class-level effect SD	9.74	11.55	16.04	18.47	19.54
Standard error	(2.75)	(2.65)	(2.39)	(2.64)	(2.61)
Number of schools	53	54	55	56	54
Number of classes	225	238	250	244	217
Mean pupils/school	29.9	30.0	30.0	30.1	31.9
Mean pupils/class	7.1	6.8	6.6	6.9	7.9

For illustrative purposes, in the three charts that follow, we order teachers, from least (left) to most (right) effective based on their 'overall' effectiveness in terms of their value-added estimate for their entire class (the model in the first column of Table 1). We present a value-added estimate for each teacher for each of the five quintile groups in the form of a 'stacked bar' in order to illustrate how each teacher's overall value-added derives from the value-added for each of these five pupil groups. Each teacher of course teaches classes of different compositions in terms of these quintile groups, which is not shown in these illustrative charts, for the purposes of readability. In other words, the size of quintile groups within a class vary in size and a teacher's overall effectiveness is a function of their effectiveness for each group and the size of each group which benefits from this effectiveness. Hence the length of bars in the charts that follow do not decline uniformly from left to right when teachers are ordered from less to more effective based on overall effectiveness. Teacher value-added estimates are reported in the test-score metric, where the standard deviation of pupil test scores is 100 at the beginning of the school year. Value-added estimates are centred on zero, so that a teacher of 'average effectiveness' has a value-added estimate of zero. The estimated mean level of pupil progress in mathematics over the school year (Grade 7 or 8) was 31 points or 0.31 standard deviations (see Rossiter et al., 2017: 35). Accordingly, we may consider the average annual learning gain to be around one third of a standard deviation on this test-score scale. If an effective teacher is considered to be one whose value-added is one standard deviation above the mean as considered above, the additional 'value-added' of such a teacher at around 0.2 standard deviations is roughly equivalent to two thirds of a year of schooling for an average pupil; a not inconsiderable value. As we see in the charts below, however, not all pupils in a class benefit equally from teacher effectiveness.

Figures 4, 5 and 6 we contrast the 'least effective' teachers with teachers of 'average effectiveness' and with the 'most effective' (ordered by their overall value-added estimates and showing their estimates for each quintile). It is clear that there is a lot of variation in both overall effectiveness and in whom teachers are effective for. This is most obvious in the middle of the distribution of teacher effectiveness (Figure 5) where many teachers have negative value-added estimates for some quintiles of pupils and positive estimates for others. Some teachers are more effective for lowest performers apparently at the cost of higher performers, and vice-versa. In contrast, the least effective teachers are perhaps especially ineffective for the most able pupils, since among the 50 least effective teachers negative columns are dominated by pupil quintiles 4 and 5.

In order to reach the highest levels of average effectiveness (Figure 6), teachers would need to be effective for all or most of the quintile groups of pupils within their classes. This can be seen clearly among the very most effective teachers whose bars by quintile are somewhat more equal in size.

[6] SD – standard deviation

Among this group of teachers, there is no clear evidence of ineffective teaching for any particular group, where prior attainment is concerned except perhaps for quintile 1 (the lowest scoring pupils). The bars for quintile 1 are relatively small for almost all teachers, even at the highest levels of effectiveness, indicating that relatively little of teachers' overall effectiveness is derived from their effectiveness in raising the attainment of the lowest performing pupils, that is those whose attainment is typically furthest from curricular expectations.

Figure 4: Differential Teacher Effectiveness in Ethiopia: Least Effective Teachers

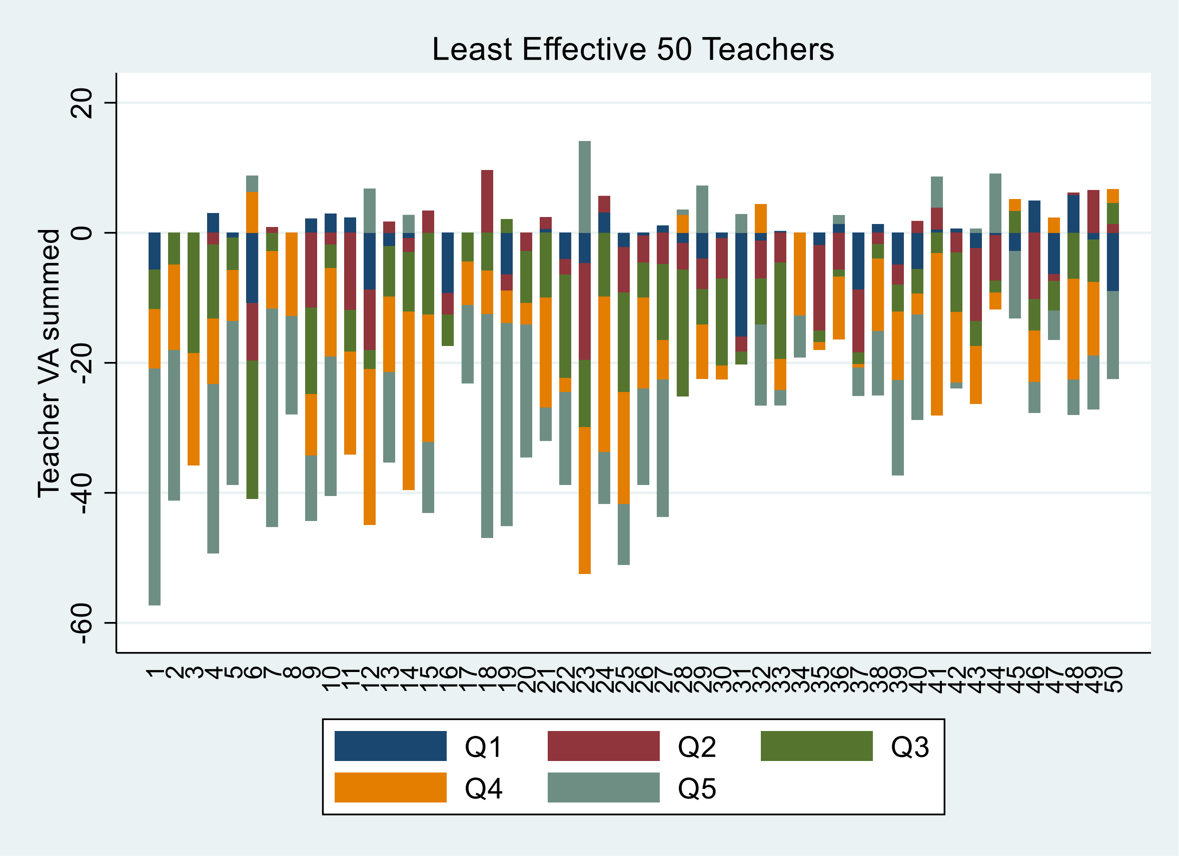


Figure 5: Differential Teacher Effectiveness in Ethiopia: Teachers of Average Effectiveness

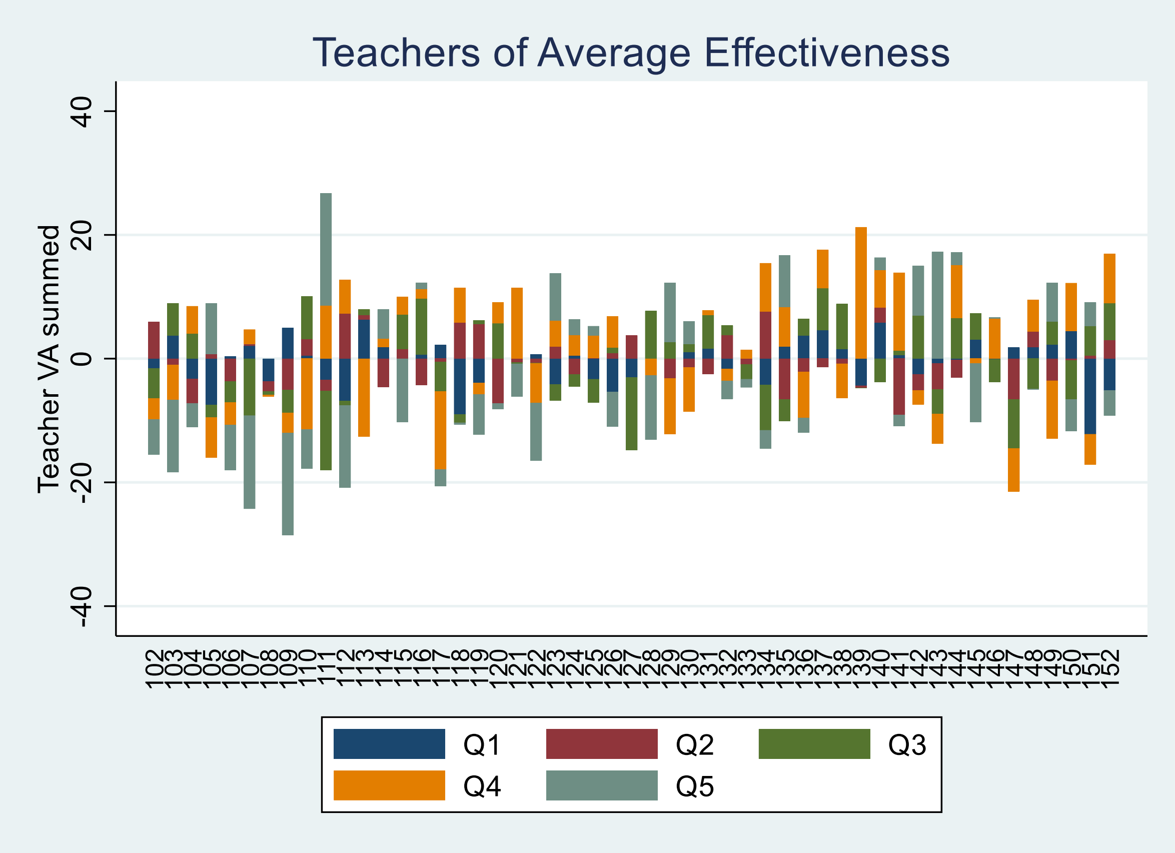
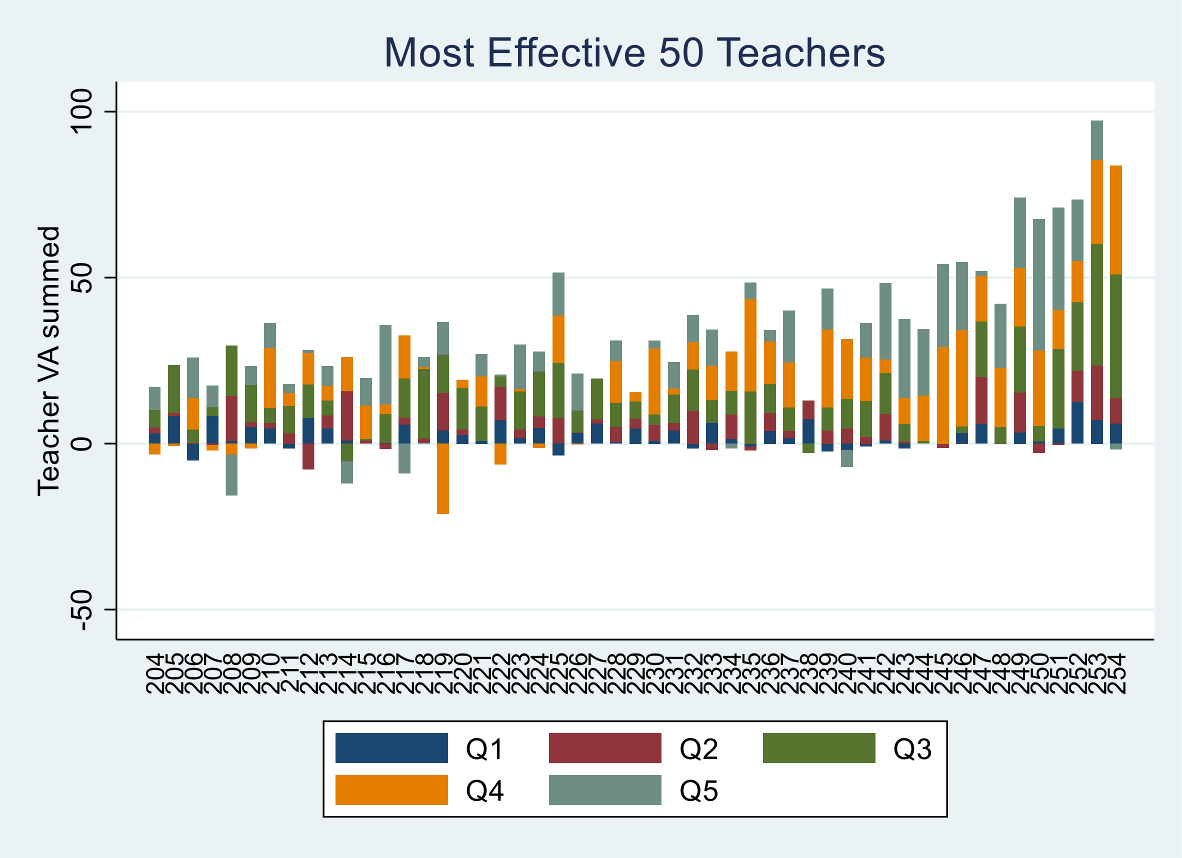


Figure 6: Differential Teacher Effectiveness in Ethiopia: Most Effective Teachers



6. Discussion and Conclusions

A number of puzzles persist in relation to empirical investigations of teacher effectiveness. In particular, while studies such as the present one show apparently large differences in overall teacher effectiveness linked to large differences in pupil progress, education production function studies typically show that observed teacher characteristics often explain little in terms of variation in pupil progress or in terms of 'black box' (residual) teacher effectiveness (see for example Bau and Das, 2020 in the context of Pakistan). School and teacher effectiveness in the empirical literature is usually estimated based on *average* progress for the relevant pupil group - class, teacher or school. It says little about differential effectiveness within classrooms or schools; and a teacher whose overall effectiveness is modest may nonetheless be highly effective for particular groups of pupils. Teachers in this study demonstrate notably differential effectiveness, as illustrated in Figures 3-6. The importance of this point has been emphasised early in the literature by Nutall et al (1989) among others:

> "School effectiveness varies in terms of the relative performance for different sub-groups. To attempt to summarize school differences, even after adjusting for intake, sex and other background characteristics of the students, in a single quantity is misleading"

A teacher's effectiveness, understood as 'value-added', depends on the composition of pupils in the classroom and on their effectiveness for each of these groups. This in turn of course depends on the teaching strategies employed and curriculum to be taught; the first being influenced by a teacher's training, support, monitoring and evaluation and so on. In classrooms where heterogeneity in performance levels is very extensive, such as in Ethiopia Grades 7 and 8 where variation in prior achievement can be equivalent to five or more grades in one classroom, differential teacher effectiveness may be of particular interest and concern. The groups benefitting least from teacher effectiveness may be expected to make little progress, and our data suggests these pupils' performance is typically already very far from curricular expectations. At the same time, we have shown that the issue of poor performance in relation to curricula is more pervasive and that even very large and wholesale improvements in teacher effectiveness by themselves would provide only a partial solution to this problem in Ethiopia.

Value-added analysis and in particular differential value-added analysis offers important potential for education system diagnostic research and practice. There is arguably even greater potential in low-income countries and for education systems affected by the 'learning crisis'. Progress in resolving the crisis will depend closely on the success of strategies to improve learning among lower performing pupils ('raising the floor'), including strategies to accelerate progress for these pupils by improving the effectiveness of their teachers, understood specifically as effectiveness for lower performers within classes and schools. Differential value-added analysis as considered in this paper may be useful for examining competing explanations of low performance as a result of weak progress or shallow learning trajectories. For example, the hypothesis that 'over-ambitious curricula' contribute to low learning outcomes and poor progress (Pritchett and Beatty, 2012), particularly for those whose learning is furthest from expectations, provides a convincing account of why learning stagnates under certain conditions. Nonetheless, while our evidence does provide empirical support for the suggestion that once pupils' learning has fallen far from expectations it may become increasingly difficult for teachers to ensure progress; we do also identify a number of cases of classrooms in which low performing pupils make good progress and benefit from effective teaching. Nutall et al (1989) highlight the potentially valuable policy implication of identifying these cases:

> "It is those schools that narrow the gap by raising the performance of the lower achieving group that may be of special interest. It would be valuable to study such schools in depth in cooperation with expert observers, such as inspectors, to explore possible reasons for their differential performance."

Education systems, both centrally and at regional and district levels, might benefit from understanding or even replicating and scaling the strategies being adopted and the adaptations being made by teachers and schools which are effective for low performers. More generally, where weak teacher and school effectiveness for low performers is found to be pervasive, there may be wider implications for curricular and pedagogical reform. In practice, the 'teaching at the right level' (TaRL) literature has explored some of these, primarily by way of experiments intended to evaluate the effects of pedagogy and instruction which focuses more closely on progress from pupils' existing levels of attainment, however low. Approaches to TaRL have, for example, included a focus on grouping pupils according to their skills (performance or attainment) rather than their age or grade-completion; and specifically on foundational skills development as well as on teacher training to support pedagogical change. Some TaRL oriented programmes have shown notable success (see Hwa, Kaffenberger and Silberstein, 2020 for a review). Our differential value-added analysis provides some indicative support for the potential of reforming grouping strategies in schools. It shows that some teachers are notably more effective at teaching pupils of different levels of prior attainment, indicating possible benefits from taking account of value-added concerns in teacher deployment and allocation, that is 'matching' of teachers to pupils who might benefit most. This might entail grouping pupils into classes which are more homogeneous with respect to prior learning (as in the TaRL approach) but clearly would involve a number of other system-wide considerations, some of which are more easily avoided in an experimental setting or NGO programme than a public education system. These include issues around teacher training and incentives. For example, where teachers are equipped specifically to teach a particular grade or incentivised to prepare pupils for a particular examination, it may be challenging to release 'untapped effectiveness' through reorganisation and reallocation alone without considerable retraining or upskilling as well as reform of incentive structures and of the mismatch between curricula and actual learning levels.

Teachers in much of sub-Saharan Africa face large and heterogenous classes in challenging and poorly resourced conditions. Such conditions are particularly demanding in terms of teacher skill and effectiveness. While teacher and school value-added analysis has been employed most commonly in high income countries in support of teacher and school evaluation and to inform parental choice; in low- and middle-income countries it may provide a valuable diagnostic tool in support of strategy and reform. Differential value-added analysis does not provide for the identification of causal pathways but can highlight strengths and weaknesses of teachers, schools and systems in respect of pupil progress and of the effectiveness with which progress is supported by schools and teachers; in turn drawing attention to avenues for potential intervention.

References

Armor, D. J., Conry-Oseguera, P., Cox, M., King, N. J., McDonnell, L. M., Pascal, A. H., Zellman, G. L. (1976). Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools. Retrieved October 20, 2018, from https://www.rand.org/pubs/reports/R2007.html

Aurino, E., James, Z., & Rolleston, C. (2014). Young Lives Ethiopia School Survey 2012–13. Young Lives, Oxford

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, *62*, 48–65. https://doi.org/10.1016/j.econedurev.2017.10.004

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Pupil Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65. https://doi.org/10.3102/10769986029001037

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., & Walton, M. (2016). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India* (No. w22746). National Bureau of Economic Research.

Barnes, A. E., Zuilkowski, S. S., Mekonnen, D., & Ramos-Mattoussi, F. (2018). Improving teacher training in Ethiopia: Shifting the content and approach of pre-service teacher education. *Teaching and Teacher Education*, *70*, 1-11.

Bau, N., and Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy, 12 (1): 62-96.*

Boyden, J., and James, Z. (2014) Schooling, Childhood Poverty and International development: choices and challenges in a longitudinal study. *Oxford Review of Education* 40.1: 10-29.

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. https://doi.org/10.1257/aer.104.9.2593

Close, K., & Amrein-Beardsley, A. (2018). Learning from what doesn't work in teacher evaluation. *Phi Delta Kappan*, *100*(1), 15–19. https://doi.org/10.1177/0031721718797115

Crouch, L. and Rolleston, C. (2017). "Raising the Floor on Learning Levels: Equitable Improvement Starts with the Tail." An Insight note from the RISE Programme. Available at https://www.riseprogramme.org/publications/raising-floor-learning-levels-equitable-improvement-starts-tail

Crouch, L., Rolleston, C., & Gustafsson, M. (2020). Eliminating global learning poverty: The importance of equalities and equity. *International Journal of Educational Development*, 102250.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation?. *Educational Researcher*, *44*(2), 132-137.

Dearden, L., Miranda, A., & Rabe-Hesketh, S. (2011). Measuring School Value Added with Administrative Data: The Problem of Missing Variables*. *Fiscal Studies*, *32*(2), 263–278. https://doi.org/10.1111/j.1475-5890.2011.00136.x

Everson, K. C. (2017). Value-Added Modeling and Educational Accountability: Are We Answering the Real Questions? *Review of Educational Research*, *87*(1), 35–70. https://doi.org/10.3102/0034654316637199

Ferrão, M. E. (2012). On the stability of value added indicators. *Quality and Quantity*, *46*(2), 627–637. http://dx.doi.org/10.1007/s11135-010-9417-6

Glewwe, P., Lambert, S., & Chen, Q. (2020). Education production functions: updated evidence from developing countries. In *The Economics of Education* (pp. 183-215). Academic Press.

Goldhaber, D., & Hansen, M. (2013). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*, *80*(319), 589–612. https://doi.org/10.1111/ecca.12002

Gulosino, C. (2018). Evaluating the Tennessee Higher Education Commission's Report Card on the value-added estimates of teacher preparation programs. *Education Policy Analysis Archives*, *26*(0), 33. https://doi.org/10.14507/epaa.26.2604

Gupta, S. (1994). The development of education, printing and publishing in Ethiopia. *The International Information & Library Review*, *26*(3), 169-180. https://doi.org/10.1006/iilr.1994.1012

Hanushek, E. A. (1971). Teacher Characteristics and Gains in Pupil Achievement: Estimation Using Micro Data. *American Economic Review*, *61*(2), 280–288.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267

Hanushek, E. A., & Rivkin, S. G. (2012). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, *4*(1), 131–157. https://doi.org/10.1146/annurev-economics-080511-111001

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal*, *51*(1), 73–112. Retrieved from JSTOR.

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of Value-Added Estimates and Characteristics of Pupils with Hard-to-Predict Achievement Levels. *Statistics and Public Policy*, *3*(1), 1–10. https://doi.org/10.1080/2330443X.2016.1182878

Horoi, I., & Ost, B. (2015). Disruptive peers and the estimation of teacher value added. *Economics of Education Review*, *49*, 180–192. https://doi.org/10.1016/j.econedurev.2015.10.002

Hwa, Y., Kaffenberger, M., & Silberstein, J. (2020). Aligning Levels of Instruction with Goals and the Needs of Students (ALIGNS): Varied Approaches, Common Principles. *RISE Insight Series*, *22*.

Giffin, J., Hershberg, T. and Robertson-Craft, C. (2009) Value-Added as a Classroom Diagnostic in Theodore Hershberg and Claire Robertson-Craft, Eds., A Grand Bargain for Education. Cambridge, Mass.: Harvard Education Press,

Isenberg, E., & Hock, H. (2010). Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools. *Final Report*, 26.

Iyer, P., Rolleston, C., Rose, P., & Woldehanna, T. (2020) A rising tide of access: what consequences for equitable learning in Ethiopia?. Oxford Review of Education https://doi.org/10.17863/CAM.50021

Jerald, C. D. (2009). The Value of Value-Added Data. K-12 Policy. *Education Trust*. Available at: https://files.eric.ed.gov/fulltext/ED507719.pdf

Kupermintz, H., Shepard, L., & Linn, R. (2001). Teacher Effects as a Measure of Teacher Effectiveness: Construct Validity Considerations in TVAAS (Tennessee Value Added Assessment System).

Kurtz, M. D. (2018). Value-Added and Pupil Growth Percentile Models: What Drives Differences in Estimated Classroom Effects? *Statistics and Public Policy*, *5*(1), 1–8. https://doi.org/10.1080/2330443X.2018.1438938

McCaffrey, D. F., Lockwood, J. R., Koretz, D., M., & Hamilton, L., S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy*, *4*(4), 572–606. Retrieved from JSTOR.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children [by]*. Cambridge, Mass., Ballinger Pub. Co.

Nuttall, D. L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13*(7), 769-776.

Pritchett, L., & Beatty, A. (2012). The negative consequences of over-ambitious curricular in developing countries: Harvard Kennedy School.

Pritchett, L. and Viarengo, M. (2021). Learning Outcomes in Developing Countries: Four Hard Lessons from PISA-D. RISE Working Paper Series. 21/069. https://doi.org/10.35489/BSG-RISE-WP_2021/069Rolleston, C., James, Z., Pasquier-Doumer, L., & Tam, T. N. T. M. (2013). *Making progress: Report of the Young Lives school survey in Vietnam*. Young Lives.

Rolleston, C., & Moore, R. (2018). *Young Lives School Survey, 2016-17: Value-added Analysis in India*. Retrieved from Young Lives website: https://ora.ox.ac.uk/objects/uuid:113e8e0f-8f39-435b-b47d-ea7f6b10ff4f/download_file?file_format=pdf&safe_filename=YL-ValueAddedAnalysis-India.pdf&type_of_work=Report

Rossiter, J., Azubuike, O. B., & Rolleston, C. (2017). Young Lives School Survey, 2016–17: Evidence from Ethiopia. Young Lives, Oxford.

Rossiter, J., Woodhead, M., Rolleston, C., & Moore, R. (2018). Delivering on every child's right to basic skills. Oxford: Young Lives. Retrieved from Young Lives website: https://www.younglives.org.uk/sites/www.younglives.org.uk/files/YL-Education.pdf

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Pupil Achievement. *Quarterly Journal of Economics*, *25*(1), 175–214.

Scheerens, J. (2000): Improving school effectiveness (Fundamentals of Educational Planning No. 68). Paris: UNESCO/International Institute for Educational Planning.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., & Jimenez, E. (2016). The impact of education programmes on learning and school participation in low-and middle-income countries. London: International Initiative for Impact Evaluation (3ie).

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of pupils they serve? *Economics of Education Review*, *64*, 50–74. https://doi.org/10.1016/j.econedurev.2018.04.001

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3-F33.

USAID (2017) English for Ethiopia Teacher Training Manual Grades 5-8. Retrieved from USAID website: https://pdf.usaid.gov/pdf_docs/PA00N2KD.pdf

World Bank. (2016). *Striving for excellence: Analysis of Ethiopia national learning assessments, 2011 and 2015*. Retrieved from World Bank website: https://documents1.worldbank.org/curated/en/580961492110426813/pdf/Ethiopia-Education-PforR-PID-20170405.pdf

Acknowledgement

We are grateful to Sirin Tangpornpaiboon of University College London for her work supporting the literature review for this study.

Appendices

A1: Descriptive Statistics Q1

Variable	Obs	Mean/ Proportion	Std. Dev.	Min	Max
Maths score T2	1861	443.591	57.321	254.633	612.515
Maths score T1	1861	382.087	29.392	251.045	420.344
English score T1	1754	430.637	57.367	265.9	753.977
Age	1798	14.092	1.515	11	27
Male	1801	.442		0	1
Mother reads	1788	.537		0	1
Father reads	1784	.69		0	1
Wealth Index	1861	-.481	1.841	-4.034	2.817
Three meals	1799	.792		0	1
Attended pre-school	1783	.551		0	1

A2: Descriptive Statistics Q2

Variable	Obs	Mean/ Proportion	Std. Dev.	Min	Max
Maths score T2	1861	476.127	63.066	266.036	650.116
Maths score T1	1861	444.829	13.772	420.377	468.456
English score T1	1764	454.925	65.395	274.351	683.68
Age	1812	14.224	1.545	11	27
Male	1806	.468		0	1
Mother reads	1791	.553		0	1
Father reads	1790	.715		0	1
Wealth Index	1861	-.274	1.83	-4.034	2.817
Three meals	1809	.79		0	1
Attended pre-school	1778	.549		0	1

A3: Descriptive Statistics Q3

Variable	Obs	Mean/ Proportion	Std. Dev.	Min	Max
Maths score T2	1861	516.651	68.764	332.454	715.794
Maths score T1	1861	494.677	15.528	468.469	522.484
English score T1	1756	489.378	73.349	316.874	790.15
Age	1826	14.189	1.401	11	21
Male	1825	.479		0	1
Mother reads	1793	.565		0	1
Father reads	1802	.713		0	1
Wealth Index	1861	.025	1.702	-4.034	2.817

Three meals	1819	.83		0	1
Attended pre-school	1804	.595		0	1

A4: Descriptive Statistics Q4

Variable	Obs	Mean/ Proportion	Std. Dev.	Min	Max
Maths score T2	1861	574.801	76.025	356.999	779.605
Maths score T1	1861	553.579	19.152	522.544	589.774
English score T1	1771	535.151	82.089	293.835	806.46
Age	1845	14.321	1.46	11	25
Male	1842	.472		0	1
Mother reads	1824	.609		0	1
Father reads	1826	.752		0	1
Wealth Index	1861	.361	1.6	-4.034	2.817
Three meals	1842	.847		0	1
Attended pre-school	1827	.646		0	1

A5: Descriptive Statistics Q5

Variable	Obs	Mean/ Proportion	Std. Dev.	Min	Max
Maths score T2	1862	673.058	89.692	403.493	889.397
Maths score T1	1862	661.146	59.986	589.794	894.646
English score T1	1785	614.78	92.457	351.028	818.99
Age	1845	14.324	1.326	11	23
Male	1841	.551		0	1
Mother reads	1837	.669		0	1
Father reads	1829	.767		0	1
Wealth Index	1862	.744	1.429	-4.034	2.817
Three meals	1847	.858		0	1
Attended pre-school	1830	.711		0	1