Received: 2 April 2019 | Revised: 29 January 2020 | Accepted: 27 July 2020

DOI: 10.1111/rode.12717



WILEY

SPECIAL ISSUE ARTICLE

# Long-run effects of teachers in developing countries

**Lee Crawfurd**[1] | **Caine Rolleston**[2]

[1]Center for Global Development, London, United Kingdom of Great Britain and Northern Ireland

[2]UCL Institute of Education, 20 Bedford Way, London, WC1H0AL, United Kingdom of Great Britain and Northern Ireland

**Correspondence**
Caine Rolleston, UCL Institute of Education, London, United Kingdom of Great Britain and Northern Ireland.
Email: c.rolleston@ucl.ac.uk

**Abstract**

How persistent are teacher effects on student outcomes? In this paper we present estimates of teacher effects on long-run student outcomes from two low- and middle- income countries. We first estimate teacher value-added using the Young Lives School Survey data from Ethiopia and Vietnam. We then track students taught by these teachers 2 and 5 years later and use data from the Young Lives Household Surveys to estimate the effects of teacher quality. We find no persistent effect after 2 years, but better mathematics ($0.08\sigma$) and reading ($0.06\sigma$) test scores after 5 years, from being taught by a $1\sigma$ better Grade 5 teacher. We find no persistent effects of good teachers on measures of more "generalized" cognitive ability, aspirations, well-being, or "grit."

**KEYWORDS**

Ethiopia, learning, teachers, value-added, Vietnam

## 1 | INTRODUCTION

How persistent are teacher effects on student outcomes? A large literature has established the effects of individual teachers on students in the United States (Hanushek & Rivkin, 2010; Koedel, Mihaly, & Rockoff, 2015). Fewer studies consider teachers in developing countries, and none in relation to longer-run outcomes. Estimates from the USA put the economic value of an effective teacher in the hundreds of thousands of dollars, based on future wage gains of their students (Chetty, Friedman, & Rockoff, 2014b). In this paper we estimate the persistent effect of teacher quality in Ethiopia and Vietnam.

Different approaches have been taken to defining and measuring teacher quality. The "Measures of Effective Teaching" (MET) project in the USA directly compared three measures: using growth in student test scores, using lesson observations, and using student feedback. The project found that the best

composite measure is based primarily on student test scores (value-added), with a small weight for lesson observation and student feedback (Kane, McCaffrey, Miller, & Staiger, 2013). Lesson observations might, however, also help to understand *why* some teachers are more effective than others, by identifying specific practices that are associated with improved outcomes (see, for example, Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016; Bruns, Gregorio, & Taut, 2016; Carnoy, Ngware, & Oketch, 2015; Ngware, Ciera, Musyoka, & Oketch, 2015).

Several studies have tested the robustness of observational estimates of teacher quality using value-added models, by comparing them with experimental or quasi-experimental estimates. These studies generally show that value-added models using observational data can provide unbiased estimates of experimentally retrieved teacher effects. This includes studies in the USA (Bacher-Hicks, Kane & Staiger, 2014; Chetty et al., 2014b; Kane et al., 2013; Kane & Staiger, 2008) and in two developing countries (Araujo et al., 2016; Buhl-Wiggers, Kerwin, Smith, & Thornton, 2017).

The longitudinal test score data needed for estimating teacher value-added models are relatively scarce in developing countries. Such data are more common in high-income countries that have extensive official digitalized testing infrastructure. The few papers focused on developing countries have instead made use of survey data (Araujo et al., 2016; Bau & Das, 2020; Buhl-Wiggers et al., 2017) and administrative data from a private school network in India (Azam & Kingdon, 2015).

Having established substantial variation in teacher quality, other studies in high-income contexts have estimated the persistence of these effects. These find causal effects of high school teachers on college attendance and earnings as an adult (Chetty, Friedman, & Rockoff, 2014a; Jackson 2016), as well as effects of primary school teachers on tertiary enrollment, employment, and earnings at age 20 (Flèche, 2017).

Finally, another relevant body of literature focuses on "matching" between teachers and students and the related notion of "differential teacher effectiveness." Teachers can be more or less effective at teaching students of their own gender, caste, and religion in India (Muralidharan & Sheth, 2016; Rawal & Kingdon, 2010) and Pakistan (Karachiwalla, 2019). Teacher's expectations of students also vary systematically, with, for example, non-black teachers having lower expectations of black students in the USA (Gershenson, Holt, & Papageorge, 2016).

Methodologically our paper is similar to Glewwe, Krutikova, and Rolleston (2017) who combine the same school survey data with preceding rounds of the longitudinal household survey data set (Rounds 2 and 3) for Vietnam and Peru. They examine how school quality affects the learning of different students (comparing groups based on a number of axes of disadvantage). Our paper differs by focusing on teacher rather than school effects, and by looking at how students perform after having left the classroom (and moved to a different school). We focus on Vietnam and Ethiopia, where the available data suit this purpose.

In this paper we test whether having a more effective teacher at age 10 (in Grade 4 or 5) affects test scores 5 years later, when students are aged 15. We find a significant effect in mathematics and reading at age 15. In Vietnam effects are larger for wealthier students in both mathematics and reading (with no difference for Ethiopia). We find no effect on more "generalized" cognitive ability measures, or on non-cognitive measures in the form of aspirations, well-being, and "grit".

The rest of the paper is organized as follows. Section 2 provides a brief discussion of relevant features of the contexts of Ethiopia and Vietnam. Section 3 describes our data, Section 4 the methods we use, and Section 5 our results. Section 6 concludes.

## 2 | CONTEXTS

Ethiopia is a low-income country and Vietnam is a lower-middle-income country, but their education outcomes are very different. Learning outcomes in Ethiopia are typical of a low-income country,

while Vietnam compares favorably with high-income countries (see Figure 1). Vietnam ranks 27th of 157 countries on the World Bank harmonized learning outcome scale, with Ethiopia ranking 131st. In science, Vietnam ranked 8th of 72 mostly high-income countries in the 2015 Programme for International Student Assessment (PISA). Results from school-based tests such as PISA are biased estimates of population averages in lower-income countries due to higher dropout rates than in OECD countries. Nonetheless, Vietnam also performs well on the component of the Young Lives household-survey-based assessment that links to the international Trends in International Mathematics and Science Study (TIMSS) scale (Singh, 2014). Vietnam appears to have avoided a "quality–quantity trade-off" in expanding education provision (Rolleston, 2016), achieving high levels of learning for a majority of pupils and also low levels of inequality in test scores. There are also a relatively high number of high-scoring pupils from disadvantaged backgrounds (see OECD, 2016). These features point towards a context in which school and teacher effectiveness are potentially particularly informative.

## 3 | DATA

We use two sources of overlapping data from the Young Lives study. The first is a longitudinal study following the same households for five survey rounds over 15 years (the "household survey"). The second is a school-based study that examines the children from the household survey alongside their

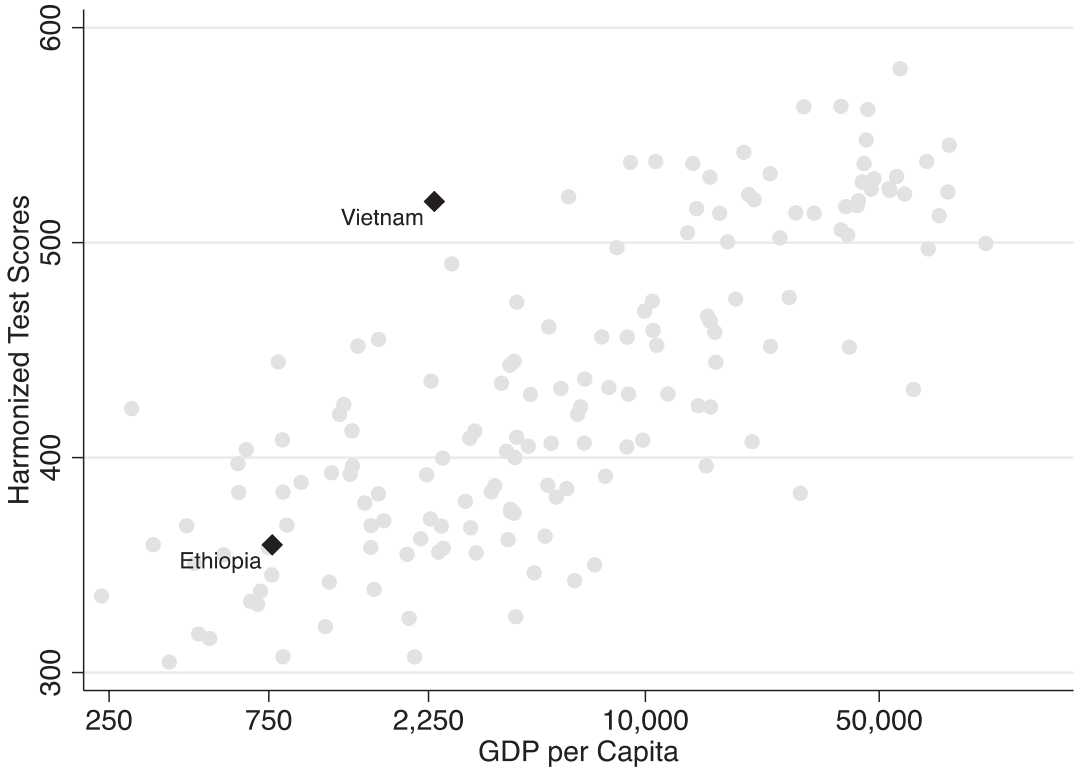


**FIGURE 1** GDP per capita and harmonized test scores. This figure presents results from the World Bank Harmonized Test Score database (Patrinos & Angrist, 2018)

**TABLE 1** Data structure

| Year | Grade | Surveys | Mean student age |
|---|---|---|---|
| 2009–10 | 3 | Student Panel Round 3 | 8 |
| 2010–11 | 4 | | 9 |
| 2011–12 | 5 | School Survey (Vietnam) | 10 |
| 2012–13 | 6 | Student Panel Round 4; School Survey (Ethiopia) | 11 |
| 2013–14 | 7 | | 12 |
| 2014–15 | 8 | | 13 |
| 2015–16 | 9 | | 14 |
| 2016–17 | 10 | Student Panel Round 5 | 15 |

classmates and teachers over a single school year (the "school survey"). The overlap in the timing of the surveys is presented in Table 1.

The household survey has followed and administered questionnaires and assessments to 12,000 children over 15 years. The survey is divided into two birth cohorts. We focus on the younger cohort (born in 2001–2), many of whom are also observed in the school survey. We make use of Rounds 3 (2009–10), 4 (2012–13), and 5 (2016–17) of the household survey. The cohort is aged 8 in Round 3, and 15 in Round 5. Each household survey includes assessments in mathematics and reading. These assessments are not specific to country curricula and are suitable for international comparisons.

The school survey was conducted in 2011–12 in Vietnam, when children were in Grade 5, and in 2012–13 in Ethiopia, when children were in Grade 4 or 5. In Ethiopia the school survey covered 13,725 students in 92 schools, of whom 549 are followed in the household survey. This includes all students in the class of the household survey child. In Vietnam the survey covered 3,284 children in 56 schools, of whom 1,138 were followed by the household survey. In Vietnam this comprised a random sample of 20 students per class (including the student from the household survey). The school survey included assessments in mathematics and reading comprehension conducted at both the beginning and end of the school year. Further details are provided in Rolleston, James, Pasquier-Doumer, and Tran (2013).

In both countries, a sentinel-site sampling design is employed, comprising 20 purposively selected sites chosen to represent national diversity, but with a pro-poor bias. At the site level, children were selected randomly in 2001 to be representative of the birth cohort in each site (see Boyden & James, 2014, for full details). The sites in Ethiopia are located in five regions; Addis Ababa, Amhara, Oromiya, Tigray, and the Southern Nations, Nationalities, and People's Region (SNNP). The sites in Vietnam are in five provinces: Lao Cai, Hung Yen, Da Nang, Phu Yen, and Ben Tre. Each province sample contains four sites, and each site is formed of one or two *woredas* in Ethiopia and *communes* in Vietnam.

Our main outcome variables are the mathematics and reading comprehension test scores of students in the Round 5 household survey (2016–17), administered during the household visits. That is 5 years after exposure to the "treatment" teacher.[1] We also consider test scores in the Round 4 household survey (2012–13), 2 years after exposure. From both Round 4 and Round 5 we also include the Peabody Picture Vocabulary Test (PPVT), a measure of receptive vocabulary which is sometimes used as a more general cognitive development indicator. In this test, the interviewer presents to the child a

series of pages that contain four pictures. The interviewer says a word and the child has to correctly identify the picture that best corresponds to the word. Further, we also employ a set of simple measures of non-cognitive skills, attitudes, or dispositions. Kraft (2019) shows that teachers have effects on both cognitive and non-cognitive skills, and Jackson (2018) shows the effects on non-cognitive skills matter more for long-run outcomes than teacher effects. We add to this literature by measuring the long-run effects on student non-cognitive skills to test whether short-run effects are persistent. Subjective well-being is measured using a Cantril ladder in which students place themselves on a scale from 1 (worst) to 9 (best). "Grit" is measured as the mean response to four questions, each answered on a four-point Likert scale. These are:

1. I can always manage to solve difficult problems if I try hard enough.
2. It is easy for me to stick to my aims and accomplish my goals.
3. I can usually handle whatever comes my way.
4. I can solve most problems if I invest the necessary effort.

To measure aspirations, we define a binary indicator for whether a student reported that they would like to complete university if they had no constraints. Finally, for Round 5 only, we calculate expected earnings at age 25, based on the average of what children reported as their expected maximum and minimum earnings at that age.

Our main treatment variable is the effectiveness of the teacher assigned in Grade 5, which is estimated using the test score data from the school survey. These test scores are derived from tests administered in schools which are in the form of multiple-choice tests specific to the country curriculum in mathematics and reading comprehension.

We also have student test scores from before they entered the treatment teacher's classroom, along with other characteristics, from the Round 3 household survey data (2009). These comprise PPVT results and scores from a basic math test appropriate for age 8.[2] Table 2 summarizes the student-level variables employed in the analysis.

The Vietnam school survey data includes 176 teachers. We drop 19 teachers from schools that have only one Grade 5 class, so that we can focus on within-school variation in teacher quality in order to limit potential biases arising from non-random selection of teachers into schools. The Ethiopia school survey includes 146 teachers. Teacher characteristics are summarized in Table 3.

## 4 | METHODS

To estimate the persistent effect of teacher quality we carry out a two-step procedure. We first estimate teacher quality using the school survey data that include test data for students at the start and end of a school year. We then link these estimates of teacher quality to later student performance.

### 4.1 | Estimating teacher quality

We estimate teacher quality with a standard student learning production function, following Todd and Wolpin (2003) and Chetty et al. (2014a). To estimate teacher effects using the school survey, we model student learning outcomes as a function of their (unobserved) ability, and all present and past individual, family, and school inputs. Lagged test scores act as a summary proxy indicator for all

**TABLE 2** Student characteristics (household survey)

| | Ethiopia | | | Vietnam | | |
|---|---|---|---|---|---|---|
| | **Mean** | ***SD*** | ***N*** | **Mean** | ***SD*** | ***N*** |
| Female | 0.50 | 0.50 | 319 | 0.48 | 0.50 | 884 |
| Age | −0.75 | 1.08 | 302 | −0.14 | 0.25 | 878 |
| Household wealth | −0.55 | 0.75 | 329 | 0.66 | 0.67 | 878 |
| PPVT score (−2 yrs) | 0.05 | 0.78 | 152 | 100.51 | 27.22 | 836 |
| Literate (−2 yrs) | 0.18 | 0.38 | 330 | 0.91 | 0.28 | 893 |
| Math score (−2 yrs) | 0.34 | 0.83 | 330 | 0.34 | 0.82 | 893 |
| Reading score (−2 yrs) | 0.18 | 0.92 | 306 | 0.12 | 0.89 | 887 |
| Math score (+5 yrs) | −0.41 | 0.68 | 318 | 0.42 | 1.00 | 880 |
| Reading score (+5 yrs) | 0.14 | 0.92 | 317 | 0.03 | 0.95 | 878 |
| PPVT score (+5 yrs) | 0.06 | 0.94 | 311 | 0.14 | 0.84 | 882 |
| Math score (+2 yrs) | −0.21 | 0.99 | 326 | 0.28 | 0.82 | 881 |
| Reading score (+2 yrs) | −0.54 | 0.73 | 323 | 0.52 | 0.81 | 882 |
| PPVT score (+2 yrs) | −0.11 | 1.01 | 316 | 0.20 | 0.68 | 890 |
| Subjective well-being (+2 yrs) | 5.88 | 1.96 | 329 | 5.55 | 1.55 | 892 |
| Subjective well-being (+5 yrs) | 5.78 | 1.73 | 319 | 6.03 | 1.52 | 884 |
| Grit (+2 yrs) | 3.16 | 1.51 | 329 | 2.91 | 0.36 | 892 |
| Grit (+5 years) | 3.17 | 0.32 | 319 | 3.18 | 1.89 | 884 |
| Aspirations (+2 yrs) | 0.69 | 0.46 | 193 | 0.75 | 0.43 | 885 |
| Aspirations (+5 yrs) | 0.79 | 0.41 | 175 | 0.70 | 0.46 | 878 |
| Exp. future income (+5 yrs) | 8.20 | 0.81 | 169 | 8.56 | 0.68 | 873 |

**TABLE 3** Teacher characteristics (school survey)

| | Ethiopia | | | Vietnam | | |
|---|---|---|---|---|---|---|
| | **Mean** | ***SD*** | ***N*** | **Mean** | ***SD*** | ***N*** |
| Female | 0.49 | 0.45 | 144 | 0.74 | 0.44 | 157 |
| Household wealth | 0.00 | 1.00 | 144 | 0.00 | 1.00 | 154 |
| Experience (yrs) | 12.45 | 8.58 | 144 | 17.63 | 8.33 | 156 |
| Higher education | 0.54 | 0.46 | 146 | 0.95 | 0.22 | 157 |
| Math specialization | 0.29 | 0.38 | 144 | 0.24 | 0.43 | 157 |
| Self-confidence | 0.00 | 1.00 | 141 | 0.00 | 1.00 | 152 |
| Permanent contract | 0.91 | 0.29 | 146 | 0.98 | 0.14 | 157 |
| Student wealth | 0.00 | 1.00 | 146 | 0.00 | 1.00 | 157 |

*Note:* Household wealth is a standardized (to mean 0 and standard deviation 1) asset index based on the first principal component of a list of assets including a phone, radio, TV, bike, car, motorbike, table, chair, fridge, electricity, and water. Higher education is a postgraduate degree in Vietnam and a post-secondary diploma or higher in Ethiopia. Self-confidence is a standardized (to mean 0 and standard deviation 1) index based on the first principal component of responses to a series of statements on self-efficacy, such as "I can get through to the most difficult students" and "I can get students to work well together". Student wealth is a standardized (to mean 0 and standard deviation 1) index of the average socioeconomic status of all individual students in their class (each being based on the first principal component of a list of household assets).

observed and unobserved inputs up to the point of that test so that we use a "value-added" framework. Moreover, by additionally controlling for contemporaneous household inputs, we can interpret any remaining changes in scores between teachers as due to that teacher. We estimate a lagged dependent variable ordinary least squares (OLS) value-added model given by

$$y_{ist} = \beta_1 \mathbf{y}_{i,t-1} + \beta_2 \mathbf{X}_{it} + \beta_3 S_s + \mu_j + \varepsilon_{ist}. \quad (1)$$

Student test scores at the end of the school year $y_{it}$ are regressed on lagged test scores[3] from the start of the school year ($y_{i,t-1}$), a rich set of student characteristics ($\mathbf{X}_{it}$) (age, gender, asset index, ethnicity, boarding status, and number of meals eaten per day), dummy variables for individual teachers $(\mu_j)$, and school fixed effects ($S_s$). As these survey-based data sets only include one cohort of children, it is impossible to distinguish between classroom and teacher effects.

Causal interpretation of estimated teacher effects $\mu_j$ (teacher quality) is impeded by the possibility of reverse causality: where there is more than one classroom per grade in a school, the allocation of students may be made by ability or aptitude ("tracking," "streaming" or "setting"). This does not appear to be a substantial issue in our data, however, since only a very small percentage of teachers report that classrooms are grouped by ability,[4] with the majority being allocated quasi-randomly. Allocation of teachers to groups of different abilities might also be non-random. Teachers might also choose particular teaching methods because of their students' abilities. However, controlling for lagged achievement should deal with this worry substantially. The "value-added" framework allows us to interpret results in terms of effects on student progress over a single school year, conditional on their starting points. Test scores from the beginning and end of the school year are calibrated concurrently using models based on item response theory so that they are directly comparable and reported on the same scale. This is possible given that the two tests contain a number of common (anchor or link) items (see Rolleston et al., 2013).

A range of specifications have been used in estimating teacher value-added in the literature. This includes two principal approaches: first, including a full set of teacher dummy variables; and second, a two-stage procedure which estimates the regression model at the student level without teacher dummies, and then averages student residuals by teacher. We prefer the full dummy set approach as it explicitly partials out any student-level covariates from the teacher effects, which controls to some extent for non-random placement of students into classrooms (demonstrated by Guarino, Reckase, Stacy, & Wooldridge, 2015, using simulated and real data).

Sampling variation in teacher effects is taken into account through adjustment based on the Bayesian shrinkage estimator (Aaronson, Barrow, & Sander, 2007). This "shrinks" estimates towards zero for teachers with small numbers of students, who would otherwise be more likely to have extreme values.

## 4.2 | Assessing sorting of students

An important concern in the estimation of teacher effects with Equation 1 is whether the systematic sorting of students of different ability into different classrooms might bias these estimates. The Rothstein falsification test shows that teacher effectiveness can be shown to predict prior student performance, which they cannot possibly causally affect (Rothstein, 2010). We replicate this finding in Table A2 in the Appendix. Several papers have argued that this test does not in fact falsify teacher value-added estimates. Goldhaber and Chaplin (2015) show theoretically and empirically

that the Rothstein test can in fact "falsify" models that are unbiased, and that the sorting of students does not necessarily imply that models are biased. Koedel and Betts (2010) show that including sufficient controls (multiple cohorts of students, and student fixed effects) can remove sorting bias, and that this bias has a relatively small effect on the estimated variation in teacher effects. The Young Lives school survey asks directly how students are assigned to classes; the overwhelming majority of teachers report that they are assigned effectively at random (and not explicitly sorted based on ability). We conduct two additional checks that do show some evidence of sorting, results of which are reported in Table A3. First, we show that, within schools, pre-existing student characteristics are only slightly correlated with teacher characteristics (after controlling for school fixed effects). In particular, students with better prior test scores are slightly more likely to be assigned to teachers with more education, though this effect is not large. We also follow the approach of Aaronson et al. (2007) and calculate the average variation (standard deviation) of test scores within classrooms. We then compare the observed variation with the variation that would be produced through perfect sorting based on prior test scores, and through random matching of students and teachers. The average standard deviation in our data is 0.78, which is closer to what would be found through perfect sorting (0.81) than through perfect random assignment (1.04). Ultimately, as we control for prior test scores we explicitly control for sorting based on these observed characteristics. Interpreting our estimates as causal requires the assumption that there is no further sorting based on unobservable characteristics. In the case where there is still sorting our estimates would be biased upwards and can hence be interpreted as an upper bound.

## 4.3 | Estimating the persistent effects of teachers

Next, we take our estimates of teacher quality with the school survey from Equation 1 and use them to predict future test scores on the household survey. We regress later test scores ($y_{i,t+1}$) on teacher quality estimated from Equation 1 ($\widehat{\boldsymbol{\mu}}_{jt}$), earlier student test scores ($y_{i,t-2}$) and covariates ($\mathbf{X}_{i,t-2}$), using OLS:

$$y_{i,t+1} = \theta_1 \widehat{\boldsymbol{\mu}}_{jt} + \theta_2 \mathbf{X}_{i,t-2} + \theta_3 \mathbf{y}_{i,t-2} + \varepsilon_{i,t+1}. \tag{2}$$

Standard errors are clustered at the teacher level. The inclusion of school fixed effects in Equation 1 reduces the chance that our results are driven by non-random sorting of teachers and students into different schools. Hence our estimates of teacher quality are based only on within-school variation in teacher quality. This means, however, that we are understating the true variation in teacher quality, which will vary across schools as well as within them.

## 4.4 | Correlates of good teachers

Finally, we estimate the correlates of teacher effectiveness:

$$\widehat{\mu}_{jt} = \gamma_1 \cdot \mathbf{Z}_{jt} + \gamma_2 \cdot \text{effort}_{jt} + \gamma_3 \cdot \text{knowledge}_{jt} + \gamma_4 \cdot \text{skill}_{jt} + \eta_{jt}. \tag{3}$$

We use the estimates of teacher quality $\widehat{\mu}_{jt}$ obtained from Equation 1 as the outcome variable, and teacher characteristics taken from the teacher interview component of the school survey.

**TABLE 4** Variation in teacher effects

| | Ethiopia | | Vietnam | |
|---|---|---|---|---|
| | No school fixed effects | School fixed effects | No school fixed effects | School fixed effects |
| 10th percentile | −0.637 | −0.728 | 0.437 | −0.005 |
| 25th percentile | −0.399 | −0.522 | 0.765 | 0.291 |
| 50th percentile | −0.190 | −0.326 | 1.034 | 0.531 |
| 75th percentile | 0.016 | −0.104 | 1.446 | 0.692 |
| 90th percentile | 0.292 | 0.122 | 1.810 | 1.014 |
| Standard deviation of TFE | 0.400 | 0.380 | 0.541 | 0.406 |
| Adjusted SD of TFE | 0.277 | 0.298 | 0.453 | 0.282 |
| $R^2$ | .649 | .564 | .574 | .457 |
| $p$-value for $F$ test of TFEs | .000 | .000 | .000 | .000 |
| Student observations | 841 | 841 | 2,060 | 2,060 |
| Teacher observations | 119 | 119 | 152 | 152 |

*Note:* This table shows the distribution of estimated teacher fixed effects (TFEs), first without school fixed effects and then with school fixed effects. The adjusted standard deviation of TFEs uses the procedure outlined in Aaronson et al. (2007) to account for the uncertainty in our estimates of the teacher effects.

An alternative approach would have been to look at the effects of teacher characteristics on student learning directly. While this might give similar coefficient estimates, we are also interested in the predictive power ($R^2$) of teacher characteristics on just the component of student learning that is affected by teachers, which is what we are estimating here.

## 5 | RESULTS

First, we estimate teacher effects using OLS (Equation 1). The standard deviation of Grade 5 teacher effects is 0.298 for Ethiopia and 0.282 for Vietnam, after applying a Bayesian shrinkage factor and controlling for school fixed effects (see Table 4). They are found to be in line with previous findings from similar contexts in India, Ecuador, and Uganda. These estimates are reported in Table A1.

We then present the results of the second-stage OLS regression of later student test scores on earlier teacher effectiveness (Equation 2). A 1 standard deviation ($\sigma$) increase in teacher quality results in an overall improvement after 5 years of $0.08\sigma$ in the pooled sample, $0.09\sigma$ in Vietnam and $0.18\sigma$ in Ethiopia. Overall these effects are around a quarter of the size of the immediate 1-year teacher effects. In reading, a $1\sigma$ increase in teacher quality results in an improvement after five years of $0.06\sigma$ in the pooled sample, $0.05\sigma$ in Vietnam and $0.11\sigma$ in Ethiopia. However, the results for the two sub-samples are not statistically significant (see Tables 5 and 6 for the results by country). We also show the results graphically in Figures 2 and A1. One possible explanation for a greater influence of (previous) teachers in math than in reading is that math is typically mostly learned at school, whereas reading is often learned in a wider range of settings including the home and broader literature environment. Table 7 reports the results over a period of 2 years and finds

**TABLE 5** Effect of teacher quality on test scores after 5 years

| | Math | Math | Math | Reading | Reading | Reading |
|---|---|---|---|---|---|---|
| Grade 5 teacher VA | 0.152*** (0.051) | 0.064** (0.032) | 0.080** (0.032) | 0.121*** (0.036) | 0.051* (0.028) | 0.063** (0.028) |
| Student controls | | Yes | Yes | | Yes | Yes |
| School FE | | | Yes | | | Yes |
| *N* (students) | 1,094 | 884 | 884 | 1,091 | 882 | 882 |
| *N* (teachers) | 256 | 202 | 202 | 256 | 202 | 202 |
| *N* (schools) | 114 | 114 | 114 | 114 | 114 | 114 |
| $R^2$ | .024 | .314 | .317 | .017 | .199 | .200 |

*Note:* This table presents a regression of 2017 student test scores on their teacher quality (value-added, VA) from 5 years prior. The dependent variable in each model is the *z*-score of the percentage of items correct on that test in the Round 5 survey. Teacher VA is estimated using the school survey data and refers to the estimated effect of the teacher on class test scores at that time (i.e. 5 years before the Round 5 test). Student controls in this regression include prior math, reading, and Peabody Picture Vocabulary Test scores (prior to exposure to the teacher in question), prior household wealth, sex, and ethnic group.

*$p < .1$; **$p < .05$; ***$p < .01$.

no significant effects in either math or reading when including student controls and school fixed effects.

## 5.1 | Heterogeneous effects (differential teacher effectiveness)

Here we examine interactions between teacher quality and baseline student characteristics (see Table 8). Students from wealthy households gain more from better teachers in Vietnam. We do not see statistically significant effects in Ethiopia, though this may be due to a smaller sample. This is in line with Glewwe et al. (2017) who find differential effects for wealthy students in Peru (but not Vietnam) over a shorter time period. More disadvantaged pupils may benefit less from teaching quality, for example, owing to a lack of social or cultural capital required to fully access the curriculum or to benefit from the pedagogical approach. This may be particularly the case for linguistic minorities or groups with large cultural differences from a dominant majority. In our results, students from wealthier backgrounds both perform better on average regardless of which teacher they had, but also benefit more than average from having been previously assigned to a quality teacher (see Table 7 for results after 2 years). Students who are $1\sigma$ above average on the household asset index benefit by $0.09\sigma$ more from having a quality teacher 5 years earlier than other students. The effect is similar in both mathematics and reading. While it is not straightforward to identify the precise channels through which these effects might operate, linked to the points above, one possibility is that wealthier households are able to provide better ongoing support to education which may help to "sustain" the benefits of good teaching over the longer term. Another possibility might relate to forms of discrimination, whether intended or not, against disadvantaged students by teachers and schools.

Looking at gender, there is no differential effect of having a high-quality teacher for boys or girls. This finding is consistent with the general pattern of high levels of gender equity in educational achievement and progress in Vietnam. In Ethiopia, while gender parity has not been reached, gender gaps are narrowing substantially. Students who are from an ethnic minority in Vietnam perform

**TABLE 6** Effect of teacher quality on test scores after 5 years, by country

| | Math | Math | Math | Reading | Reading | Reading |
|---|---|---|---|---|---|---|
| Vietnam | | | | | | |
| Grade 5 teacher VA | 0.160** (0.062) | 0.079** (0.034) | 0.092*** (0.032) | 0.090* (0.048) | 0.009 (0.031) | 0.048 (0.029) |
| Student controls | | Yes | Yes | | Yes | Yes |
| School FE in Equation 1 | | | Yes | | | Yes |
| *N* (students) | 745 | 745 | 745 | 745 | 745 | 745 |
| *N* (teachers) | 153 | 153 | 153 | 153 | 153 | 153 |
| *N* (schools) | 73 | 73 | 73 | 73 | 73 | 73 |
| $R^2$ | .026 | .307 | .311 | .008 | .218 | .221 |
| Ethiopia | | | | | | |
| Grade 5 teacher VA | 0.374*** (0.080) | 0.304** (0.119) | 0.176* (0.097) | 0.358*** (0.070) | 0.174 (0.106) | 0.111 (0.088) |
| Student controls | | Yes | Yes | | Yes | Yes |
| School FE | | | Yes | | | Yes |
| *N* (students) | 408 | 183 | 183 | 407 | 183 | 183 |
| *N* (teachers) | 137 | 77 | 77 | 137 | 77 | 77 |
| *N* (schools) | 73 | 73 | 73 | 73 | 73 | 73 |
| $R^2$ | .085 | .267 | .244 | .087 | .175 | .166 |

*Note:* This table presents a regression of 2017 student test scores on their teacher quality (value-added, VA) from 5 years prior. The dependent variable in each model is the *z*-score of the percentage of items correct on that test in the Round 5 survey. Teacher VA is estimated using the school survey data, and refers to the estimated effect of the teacher on class test scores at that time (i.e. 5 years before the Round 5 test). Student controls in this regression include prior math, reading, and Peabody Picture Vocabulary Test scores (prior to exposure to the teacher in question), prior household wealth, sex, and ethnic group.

*$p < .1$; **$p < .05$; ***$p < .01$.

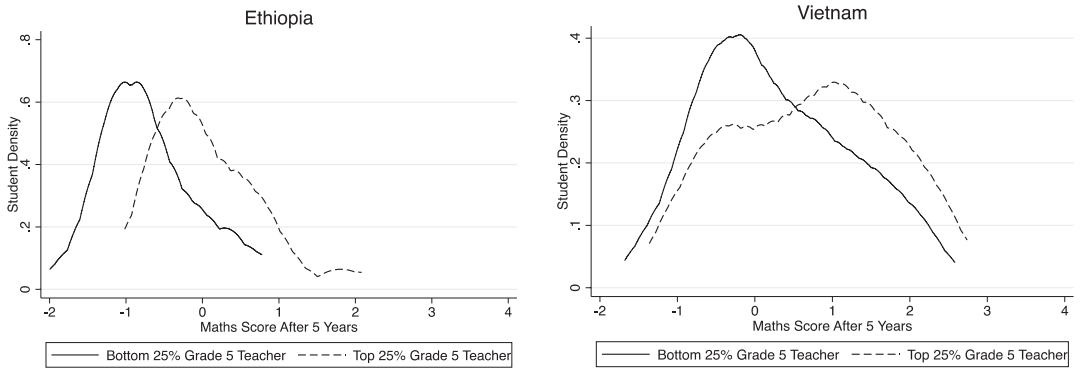


**FIGURE 2** Distribution of test scores 5 years after grade 5 teacher assignment. This figure presents the distribution of student mathematics test scores five years after Grade 5, for students with a bottom quartile Grade 5 teacher, and those with a top quartile Grade 5 teacher

**TABLE 7** Effects of teacher quality on test scores after 2 years

| | Math | Math | Math | Reading | Reading | Reading |
|---|---|---|---|---|---|---|
| Teacher VA | 0.129*** (0.047) | 0.054* (0.031) | 0.039 (0.032) | 0.109** (0.043) | 0.041 (0.030) | 0.027 (0.032) |
| Student controls | | Yes | Yes | | Yes | Yes |
| School FE | | | Yes | | | Yes |
| *N* (students) | 1098 | 875 | 875 | 1096 | 872 | 872 |
| *N* (teachers) | 256 | 200 | 200 | 256 | 200 | 200 |
| *N* (schools) | 114 | 114 | 114 | 114 | 114 | 114 |
| $R^2$ | .023 | .279 | .276 | .015 | .297 | .296 |

*Note:* This table presents a regression of 2013 student test scores on their teacher quality (value-added, VA) from 1–2 years prior. The dependent variable in each model is the *z*-score of the percentage of items correct on that test in the Round 4 survey. Teacher VA is estimated using the school survey data, and refers to the estimated effect of the teacher on class test scores at that time (i.e. 1–2 years before the Round 4 test). Student controls in this regression include prior math, reading, and Peabody Picture Vocabulary Test scores (prior to exposure to the teacher in question), prior household wealth, sex, and ethnic group.

*$p < .1$; **$p < .05$; ***$p < .01$.

substantially worse on average, but their reading benefits more than average from having previously been assigned a quality teacher both 2 (see Table 7) and 5 years ago. Ethnic minority students, unlike majority Kinh, often do not speak Vietnamese at home, so the importance of school and teaching in their learning of Vietnamese may be expected to be greater, which is consistent with this finding. For Ethiopia there is not a clear ethnic minority group. In Amhara and Addis Ababa over 95% of our sample speak Amharic, and 100% speak Tigrinya in our Tigray sample. There is an Amharic-speaking minority in our Oromiya sample, and three minority language groups in the SNNP, but the sample of these groups is small.

We examine possible effects of effective teachers on a number of non-cognitive measures collected in the Young Lives surveys, reported in Tables 9 and 10. No significant teacher effects on these outcomes are detected.

**TABLE 8** Heterogeneous effects of teacher quality on test scores after 5 years

| | Math | Math | Math | Reading | Reading | Reading |
|---|---|---|---|---|---|---|
| Vietnam | | | | | | |
| Teacher VA | 0.101*** (0.038) | 0.090*** (0.032) | 0.087*** (0.032) | 0.038 (0.032) | 0.046 (0.029) | 0.048* (0.027) |
| Teacher VA × Girl | −0.017 (0.050) | | | 0.027 (0.046) | | |
| Teacher VA × Ethnic minority | | 0.141 (0.112) | | | 0.581*** (0.080) | |
| TVA × Wealth | | | 0.094*** (0.035) | | | 0.087*** (0.032) |
| Student controls | Yes | Yes | Yes | Yes | Yes | Yes |
| *N* (students) | 745 | 745 | 745 | 745 | 745 | 745 |
| *N* (teachers) | 153 | 153 | 153 | 153 | 153 | 153 |
| *N* (schools) | 73 | 73 | 73 | 73 | 73 | 73 |
| $R^2$ | .311 | .312 | .319 | .221 | .225 | .228 |
| | **Math** | **Math** | | **Reading** | **Reading** | |
| Ethiopia | | | | | | |
| Teacher VA | 0.178* (0.100) | 0.188** (0.086) | | 0.107 (0.085) | 0.137 (0.106) | |
| Teacher VA × Girl | 0.077 (0.183) | | | −0.110 (0.168) | | |
| TVA × Wealth | | −0.025 (0.129) | | | −0.055 (0.097) | |
| Student controls | Yes | Yes | | Yes | Yes | |
| *N* (students) | 183 | 183 | | 183 | 183 | |
| *N* (teachers) | 77 | 77 | | 77 | 77 | |
| *N* (schools) | 73 | 73 | | 73 | 73 | |
| $R^2$ | .245 | .244 | | .169 | .168 | |

*Note:* This table presents a regression of 2017 student test scores on their teacher quality (value-added, VA) from 5 years prior. The dependent variable in each model is the *z*-score of the percentage of items correct on that test in the Round 5 survey. Teacher VA is estimated using the school survey data, and refers to the estimated effect of the teacher on class test scores at that time (i.e. 5 years before the Round 5 test). Student controls include sex, ethnic group, asset index, and whether they board at school.

*$p < .1$; **$p < .05$; ***$p < .01$.

**TABLE 9** Non-cognitive outcomes

| | Happiness (+2 Years) | Happiness (+5 Years) | Grit (+2 Years) | Grit (+5 Years) |
|---|---|---|---|---|
| Grade 5 teacher VA | −0.016 (0.048) | 0.016 (0.053) | 0.003 (0.015) | 0.087 (0.058) |
| Student controls | Yes | Yes | Yes | Yes |
| *N* (students) | 887 | 888 | 886 | 888 |
| *N* (teachers) | 201 | 202 | 201 | 202 |
| *N* (schools) | 114 | 114 | 114 | 114 |
| $R^2$ | 0.019 | 0.055 | 0.024 | 0.004 |

*Note:* This table presents a regression of later student happiness and grit scores on their teacher quality (value-added, VA) from Grade 5. Teacher VA is estimated using the school survey data, and refers to the estimated effect of the teacher on class test scores at that time. Student controls in this regression include prior math, reading, and Peabody Picture Vocabulary Test scores (prior to exposure to the teacher in question), prior household wealth, sex, and ethnic group.

*$p < .1$; **$p < .05$; ***$p < .01$.

**TABLE 10** Aspirations

| | University aspirations (+2 Years) | University aspirations (+ 5 Years) | (Log) expected income (+5 Years) |
|---|---|---|---|
| Grade 5 teacher VA | 0.021 (0.015) | 0.008 (0.019) | −0.002 (0.022) |
| Student controls | Yes | Yes | Yes |
| *N* (students) | 820 | 817 | 809 |
| *N* (teachers) | 182 | 180 | 176 |
| *N* (schools) | 114 | 114 | 114 |
| $R^2$ | 0.112 | 0.128 | 0.061 |

*Note:* This table presents a regression of later student aspirations for further education and expected income on their teacher quality (value-added, VA) from Grade 5. Teacher VA is estimated using the school survey data, and refers to the estimated effect of the teacher on class test scores at that time. Student controls in this regression include prior math, reading, and Peabody Picture Vocabulary Test scores (prior to exposure to the teacher in question), prior household wealth, sex, and ethnic group.

*$p < .1$; **$p < .05$; ***$p < .01$.

## 5.2 | Correlates of teacher quality

We proceed to collapse the data to the teacher level in order to examine which teacher and classroom characteristics correlate with teacher effectiveness. Here we see that few characteristics of the teachers themselves are strongly correlated with performance (see Table 11). This is consistent with the literature in general, which finds that observed characteristics of teachers have typically weak explanatory power in educational production function studies (see Glewwe et al., 2020). We include teacher age, an asset index,[5] years of experience, higher education (a postgraduate degree in Vietnam and a post-secondary diploma or higher in Ethiopia), whether teachers have a math specialization, a teacher self-efficacy index,[6] and whether they are on a permanent contract. The classroom-average student asset index is correlated with teacher performance if we do not control for school fixed effects, but not if we do (indicating that there is sorting of wealthier students to good schools, but not to good teachers within schools).

**TABLE 11** What makes a good teacher?

| | Pooled | Pooled | ET | VT |
|---|---|---|---|---|
| Female | 0.224* (0.132) | 0.168 (0.136) | 0.322** (0.162) | 0.170 (0.235) |
| Asset index | −0.027 (0.046) | 0.036 (0.047) | 0.115 (0.090) | 0.010 (0.065) |
| Experience (years) | 0.003(0.007) | 0.005 (0.007) | −0.010 (0.010) | 0.023* (0.012) |
| Higher education | 0.008 (0.148) | −0.090 (0.152) | 0.070 (0.155) | 0.081 (0.444) |
| Math specialization | 0.128 (0.145) | 0.098 (0.148) | 0.005 (0.191) | 0.286 (0.228) |
| Self-confidence | 0.027 (0.059) | 0.023 (0.061) | −0.014 (0.068) | 0.050 (0.101) |
| Permanent contract | 0.364 (0.281) | 0.377 (0.288) | 0.490* (0.272) | 0.249 (0.694) |
| Student asset index | 0.194*** (0.068) | 0.003 (0.070) | −0.167 (0.106) | 0.080 (0.097) |
| School FE | | Yes | Yes | Yes |
| *N* | 280 | 280 | 133 | 147 |
| $R^2$ | .051 | .021 | .108 | .045 |

*Note:* This table presents a regression of Grade 5 teacher value added (VA) scores on teacher characteristics. The dependent variable in each model is standardized. Household wealth is a standardized (to mean 0 and standard deviation 1) asset index based on the first principal component of a list of assets including a phone, radio, TV, bike, car, motorbike, table, chair, fridge, electricity, and water. Higher education is a postgraduate degree in Vietnam and a post-secondary diploma or higher in Ethiopia. Self-confidence is a standardized (to mean 0 and standard deviation 1) index based on the first principal component of responses to a series of statements on self-efficacy, such as "I can get through to the most difficult students" and "I can get students to work well together". Student wealth is a standardized (to mean 0 and standard deviation 1) index of the average socioeconomic status of all individual students in their class (each being based on the first principal component of a list of household assets).

*$p < .1$; **$p < .05$; ***$p < .01$.

## 6 | CONCLUSION

This paper has shown that the impacts of effective teachers are persistent. Having a "better" Grade 5 teacher results in better test scores in Grade 10, after students have graduated to middle school. Effects are larger in mathematics than in reading and are not significant for non-cognitive outcomes such as grit, aspirations, and subjective well-being. Measured teacher characteristics bear little relation to estimated teacher quality.

How much is this worth in economic terms? Chetty et al. (2014b) estimate the value of an effective teacher (specifically of replacing a teacher in the bottom 5% of the value-added distribution with one of average quality in value-added terms) in the USA to be around $250,000 per classroom over a teacher's career. This is based on the increase in the group of students' lifetime incomes associated with test score improvements which are expected from more effective teaching in this illustration. Estimates of the rate of return to higher skills in Vietnam suggest that a 1 standard deviation higher reading test score is associated with 15% higher earnings overall, and 6% after controlling for schooling (Valerio, Puerta, Laura, Monroy Taborda, & Tognatta, 2016). Our results indicate that having a 1 standard deviation better teacher is associated with persistent 0.1 standard deviation better test scores, which should therefore equate to 0.6% higher earnings. Gross domestic product per capita in Vietnam is around $6,000, so a 0.6% increase is worth $36 per year per student. Each class of 20 students taught by a better teacher might therefore be expected to earn a cumulative total of $720 in addition per year. The benefits potentially go far beyond those directly evaluable in economic terms, of course.

Improved educational outcomes are associated with a very wide range of social benefits from reduced fertility to improved civic participation.

Potential policy implications are numerous, while it is not possible to "read off" implications given the uncertainty about the mechanisms involved. Nonetheless, evidence on wide variation in teacher effectiveness and on the long-lasting impacts of teacher effectiveness points towards the need for careful attention to effectiveness concerns when recruiting, training, deploying, rewarding, promoting, and managing teachers. Policies which improve teacher effectiveness at scale have the potential to bring extensive benefits, while these are likely to be somewhat context dependent.

One possible policy implication is that the importance of within-school variation in teacher quality raises questions about the potential for parental choice to lead to school improvement. Even if parents choose schools, they do not choose teachers. More broadly, our results re-emphasize a neglected point about important and consequential differences in teacher effectiveness that are not well captured by a teacher's level of qualification. The results therefore have implications for how governments should think about recruitment and performance management. However, further details on such policies are beyond the scope of this paper.

## CONFLICT OF INTEREST

The authors certify that they have no conflicts of interests arising in relation to this paper, its subject matter or materials discussed herein.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available via the UK Data Service at: beta.ukdataservice.ac.uk/datacatalogue/series/series?id = 2000060, https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000060 data series numbers 7931-2, 7823-1, 7663-1, 8357-1, 6583-1.

## ORCID

*Lee Crawfurd* https://orcid.org/0000-0003-1513-934X
*Caine Rolleston* https://orcid.org/0000-0002-4564-4331

## ENDNOTES

[1] Mathematics and reading comprehension tests comprise sets of multiple-choice questions (in the case of reading comprehension a set of questions for each of several reading texts) grouped by difficulty level. A group of sets of questions in each subject is administered in all countries, while some sets are administered only in particular countries (the most difficult in Vietnam and the least difficult in Ethiopia, for example). Questions are selected and adapted to be valid for all contexts in the study countries and to be appropriate given school curricula and the diversity of settings. Tests are administered as far as possible in local languages. In order to provide for direct comparability, a number of sets of questions were retained in Round 5 from Round 4, while overall difficulty is increased by adding a greater number of more difficult sets. In the school tests used to estimate teacher value-added the overlap between beginning- and end-of-year tests is relatively high at around 40–60% of items, but being higher in Ethiopia than in Vietnam. The results are calibrated on an item response theory scale and the time-lapse between tests is short (less than 1 year), so we consider these scores highly comparable. With regard to household tests, overlap varies by country. For example, pupils typically make more progress in Vietnam, so tests differ more over time. Nonetheless, between Rounds 4 and 5, typically up to 40% of items remain the same in mathematics and reading comprehension, while 3–4 years elapse between rounds, so direct comparability is less than in the school surveys. The Peabody Picture Vocabulary Test is essentially the same test at each round, but as children grow older, they reach a later stage on this test.

[2] More details on the tests for Rounds 1–3 are available in Boyden and James (2014).

[3] Lagged test scores include mathematics score, reading score, both of their quadratic terms, and an interaction term.

[4] Eight percent in Vietnam, 17% in Ethiopia.

[5] Standardized to a mean of 0 and standard deviation of 1, based on the first principal component of a list of assets including a phone, radio, TV, bike, car, motorbike, table, chair, fridge, electricity, and water.

[6] This is a standardized (to mean 0 and standard deviation 1) index based on the first principal component of responses to a series of statements on self-efficacy, such as "I can get through to the most difficult students" and "I can get students to work well together".

## REFERENCES


Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, *131*(3), 1415–1453.

Azam, M., & Kingdon, G. G. (2015). Assessing teacher quality in India. *Journal of Development Economics*, *117*, 74–83.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (NBER Working Paper. No. 20657). Cambridge, MA: National Bureau of Economic Research.

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

Boyden, J., & James, Z. (2014). Schooling, childhood poverty and international development: Choices and challenges in a longitudinal study. *Oxford Review of Education*, *40*, 10–29.

Bruns, B., Gregorio, S. D., & Taut, S. (2016). Measures of effective teaching in developing countriesResearch on Improving Systems of Education (RISE) Working Paper no. 009 (2016).

Buhl-Wiggers, J., Kerwin, J., Smith, J., & Thornton, R. (2017). *The impact of teacher effectiveness on student learning in Africa*, . Presented at the CSAE Conference 2017.Oxford : Centre for the Study of African Economies Conference.

Carnoy, M., Ngware, M., & Oketch, M. (2015). The role of classroom resources and national educational context in student learning gains: Comparing Botswana, Kenya, and South Africa. *Comparative Education Review*, *59*, 199–233.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Flèche, S. (2017). *Teacher quality, test scores and non-cognitive skills: Evidence from primary school teachers in the UK* (CEP Discussion Paper No 1472). London: Centre for Economic Performance.

de Ree, J., Muralidhara, K., Pradhan, M., & Rogers, H. (2014). Double for Nothing? The Effects of Unconditional Teacher Salary Increases on Student Performance, Conference Paper CESIfo Economics of Education Meeting, Munich.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209–224.

Glewwe, P., Krutikova, S., & Rolleston, C. (2017). Do schools reinforce or reduce learning gaps between advantaged and disadvantaged students? Evidence from Vietnam and Peru. *Economic Development and Cultural Change*, *65*(4), 699–739.

Glewwe, Paul, Sylvie, Lambert, & Qihui, Chen (2020). Education production functions: updated evidence from developing countries, The Economics of Education (183–215). Academic Press.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein falsification test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*, 8–34.

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015). A comparison of student growth percentile and value-added models of teacher performance. *Statistics and Public Policy*, *2*, e1034820. Retrieved from https://doi.org/10/gftrs9

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

Jackson, C. K. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes* (NBER Working Paper No. 22226). Cambridge, MA: National Bureau of Economic Research.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

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). Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

Karachiwalla, N. (2019). A teacher unlike me: Social distance, learning, and intergenerational mobility in developing countries. *Economic Development and Cultural Change*, *67*, 225–271.

Koedel, C., & Betts, J. R. (2010). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*, 18–42.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Kraft, Matthew A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1–36.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of political Economy*, *119*(1), 39–77.

Muralidharan, K., & Sheth, K. (2016). Bridging education gender gaps in developing countries: The role of female teachers. *Journal of Human Resources*, *51*(2), 269–297.

Ngware, M. W., Ciera, J., Musyoka, P. K., & Oketch, M. (2015). Quality of teaching mathematics and learning achievement gains: Evidence from primary schools in Kenya. *Educational Studies in Mathematics*, *89*, 111–131. https://doi.org/10.1007/s10649-015-9594-2

OECD (2016). *PISA 2015 Results (Volume I) – Excellence and Equity in Education*, Paris: OECD Publishing. http://dx.doi.org/10.1787/19963777.

Patrinos, Harry Anthony, & Angrist, Noam (2018). *Global Dataset on Education Quality : A Review and Update (2000-2017) (English)*. Policy Research working paper;no. WPS 8592, Washington, D.C.: World Bank Group. http://documents.worldbank.org/curated/en/390321538076747773/Global-Dataset-on-Education-Quality-A-Review-and-Update-2000-2017.

Rawal, S., & Kingdon, G. (2010). *Akin to my teacher: Does caste, religious or gender distance between student and teacher matter? Some evidence from India* (DoQSS Working Paper No. 10–18). London: Department of Quantitative Social Science, Institute of Education, University of London.

Rolleston, C. (2016). *Escaping a low-level equilibrium of educational quality* (Working Paper No. RISE-WP-16/008). Oxford: Research on Improving Systems of Education, Blavatnik School of Government, University of Oxford.https://riseprogramme.org/research?f%5B0%5D=publication_type%3AWorking%20Paper

Rolleston, C., James, Z., Pasquier-Doumer, L., & Tran, N. T. M. T. (2013). *Making progress: Report of the Young Lives school survey in Vietnam* (Working Paper 100). Oxford: Young Lives, Oxford Department of International Development, University of Oxford.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*, 175–214.

Singh, A. (2014). *Emergence and evolution of learning gaps across countries: panel evidence from Ethiopia, India, Peru and Vietnam* (Working Paper 124). Oxford: Young Lives, Oxford Department of International Development, University of Oxford.

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

Valerio, A., Puerta, S., Laura, M., Monroy Taborda, S., & Tognatta, N. R. (2016). *Are there skills payoffs in low- and middle-income countries? Empirical evidence using STEP data* (Policy Research Working Paper No. 7879). Washington, DC: World Bank.

**How to cite this article:** Crawfurd L, Rolleston C. Long-run effects of teachers in developing countries. *Rev Dev Econ*. 2020;00:1–21. https://doi.org/10.1111/rode.12717

## APPENDIX

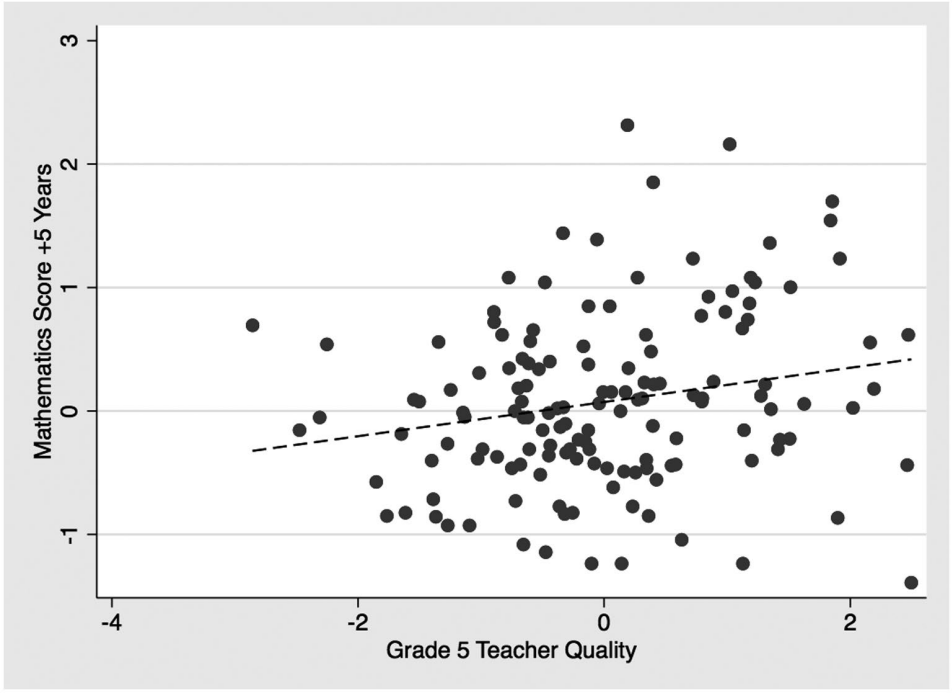


**FIGURE A1** Teacher quality and test scores after 5 years. This figure shows teacher quality estimated using *x* year school survey data, and the average mathematics scores of their students 5 years later

**TABLE A1** Distribution of teacher effects (standard deviation of student scores)

| | | Teacher Effectiveness | | |
|---|---|---|---|---|
| **Study** | **Location** | **Reading** | **Math** | **Across Subjects** |
| **United States** | | | | |
| Rockoff (2004) | New Jersey | 0.10 | 0.11 | – |
| Nye et al. (2004) | Tennessee | 0.07 | 0.13 | – |
| Rivkin et al. (2005) | Texas | 0.15 | 0.11 | – |
| Aaronson et al. (2007) | Chicago | – | 0.13 | – |
| Kane et al. (2008) | New York City | 0.08 | 0.11 | – |
| Jacob and Lefgren (2008) | Midwest city | 0.12 | 0.26 | – |
| Kane and Staiger (2008) | Los Angeles | 0.18 | 0.22 | – |
| Koedel and Betts (2011) | San Diego | – | 0.23 | – |
| Rothstein (2009) | North Carolina | 0.11 | – | – |
| Hanushek and Rivkin (2010a) | Texas city | – | 0.11 | – |
| Mean | | 0.13 | 0.17 | – |
| **Developing countries** | | | | |
| Araujo et al. (2016) | Ecuador | 0.13 | 0.11 | – |
| Azam and Kingdon (2015) | India | – | – | 0.19 |
| Buhl-Wiggers et al. (2017) | Uganda | 0.27 | – | – |
| Bau and Das (2020) | Pakistan | – | – | 0.16 |
| De Ree et al. (2014) | Indonesia | – | – | 0.35 |
| Muralidharan and Sundararam (2011) | India (AP) | – | – | 0.35 |
| Mean | | 0.20 | 0.11 | 0.26 |

*Note:* US studies are taken from Table 1 in Hanushek and Rivkin (2010).

**TABLE A2** Rothstein falsification test: Effect of teacher on prior test score growth

| | **Math** | **Math** | **Math** | **Reading** | **Reading** | **Reading** |
|---|---|---|---|---|---|---|
| Teacher VA (academics) | 0.097* (0.051) | 0.098* (0.053) | 0.104* (0.054) | 0.035 (0.045) | 0.017 (0.044) | −0.003 (0.044) |
| Student controls | | Yes | Yes | | Yes | Yes |
| School FE | | | Yes | | | Yes |
| *N* (students) | 892 | 778 | 778 | 886 | 773 | 773 |
| *N* (Teachers) | 145 | 142 | 142 | 144 | 141 | 141 |
| $R^2$ | .014 | .097 | .100 | .002 | .069 | .069 |

This table estimates the effect of Grade 5 teachers on prior growth in test scores. The outcome variable is test scores in the Round 3 longitudinal survey. Student controls include prior scores on the Peabody Picture Vocabulary Test (from Round 2 of the longitudinal survey), sex, age, and whether they are an ethnic minority.

*$p < .1$; **$p < .05$; ***$p < .01$.

**TABLE A3** Matching of students and teachers

| | Experience | Education | SES | Female | Permanent |
|---|---|---|---|---|---|
| Girl | −0.034 (0.081) | 0.001 (0.004) | −0.020* (0.011) | −0.005 (0.005) | −0.001 (0.002) |
| Student age (years) | −0.059 (0.076) | 0.014*** (0.004) | −0.009 (0.009) | 0.003 (0.004) | −0.000 (0.002) |
| Asset index | −0.059 (0.138) | −0.004 (0.007) | 0.000 (0.014) | −0.000 (0.008) | 0.000 (0.002) |
| Meals per day | −0.140 (0.109) | −0.001 (0.008) | −0.033* (0.020) | 0.025* (0.013) | 0.004 (0.005) |
| Lagged numeracy score | −0.043 (0.121) | 0.010** (0.005) | −0.024 (0.018) | −0.010 (0.006) | 0.001 (0.003) |
| Lagged literacy score | 0.034 (0.080) | 0.014** (0.005) | 0.014 (0.012) | 0.016*** (0.006) | 0.002 (0.002) |
| $F$ | 0.409 | 2.637 | 1.184 | 2.123 | 0.791 |
| $P > F$ | 0.872 | 0.018 | 0.317 | 0.053 | 0.578 |
| $N$ (students) | 28,787 | 29,034 | 28,721 | 28,837 | 29,034 |
| $N$ (teachers) | 435 | 435 | 435 | 435 | 435 |
| $N$ (schools) | 164 | 164 | 164 | 164 | 164 |
| $R^2$ | 0.000 | 0.008 | 0.002 | 0.002 | 0.000 |

In this table teacher characteristics (experience, education, socioeconomic status (SES), sex, and contract status) are regressed on prior student characteristics (sex, age, SES, meals, and lagged test scores). This shows some evidence for non-random matching of students with teachers, as students with better prior test scores are more likely to be matched with teachers with higher levels of education.

*$p < .1$; **$p < .05$; ***$p < .01$.