American Economic Journal: Economic Policy 2020, 12(1): 62–96
https://doi.org/10.1257/pol.20170243

Teacher Value Added in a Low-Income Country†

By Natalie Bau and Jishnu Das*

Using data from Pakistan, we show that existing methods produce unbiased and reliable estimates of teacher value added (TVA) despite significant differences in context. Although effective teachers increase learning substantially, observed teacher characteristics account for less than 5 percent of the variation in TVA. The first two years of tenure and content knowledge correlate with TVA in our sample. Wages for public sector teachers do not correlate with TVA, although they do in the private sector. Finally, teachers newly entering on temporary contracts with 35 percent lower wages have similar distributions of TVA to the permanent teaching workforce. (*JEL* I21, J31, J41, J45, O15)

Teacher recruitment and compensation are among the most contentious issues in education today, and in particular, an issue where research findings and policy practice diverge. For instance, policymakers frequently base recruitment on observed characteristics, but there is little empirical support for such policies. In the United States, "effective" teachers increase children's test scores and ultimately, their earnings (Chetty, Friedman, and Rockoff 2014a, 2014b), but observed teacher characteristics are only weakly related to teacher effectiveness (Rockoff 2004). Greater variation in observed teacher characteristics could justify similar recruitment policies in low-income countries, but sparse data has limited research in these contexts.

This paper addresses that gap. We examine the extent to which effective teachers increase student learning, as well as the correlation between teacher effectiveness and teacher characteristics. Our estimates are based on unique data collected between 2003 and 2007 from 112 villages in the province of Punjab, Pakistan, as part of the Learning and Educational Achievement in Pakistan Schools (LEAPS)

* Bau: UCLA, CEPR, and CIFAR, 337 Charles Young Dr. E., Los Angeles, CA 90095 (email: nbau@ucla.edu); Das: Georgetown, 37th and O Street NW, 206E Old North, Washington, DC 20057 (email: jishnu.das1@georgetown.edu). John Friedman was coeditor for this article. An earlier version of this paper was circulated as "The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers." Natalie Bau gratefully acknowledges the support of the CIFAR Azrieli Global Scholarship, National Science Foundation Graduate Research Fellowship, and Harvard Inequality and Social Policy Fellowship. Jishnu Das acknowledges funding from RISE. We thank two anonymous referees for comments that substantially improved the paper. We are also grateful to Christopher Avery, Deon Filmer, Caroline Hoxby, Asim Khwaja, Michael Kremer, Nathan Nunn, Roland Fryer, Owen Ozier, Faisal Bari, and seminar participants at the World Bank, NBER Education Meetings, CERP, NEUDC, IADB, the University of Auckland, PUC-Chile, and the University of Delaware for helpful comments. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its executive directors, or the governments they represent.

† Go to https://doi.org/10.1257/pol.20170243 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

project. These data include test score information on matched student-teacher pairs, permitting teacher value-added (TVA) estimation for 1,533 public school teachers from 574 public primary schools, who taught 22,857 unique students in grades 3, 4, and 5. We further combine these data with similar information on private school teachers and a change in the public schools' hiring regime to present new evidence on the link between compensation and TVA in the labor market for teachers.

While a number of studies have evaluated school value added in low-income countries (see Andrabi et al. 2011, Singh 2015, and Singh 2017), this is one of the first studies to estimate and validate teacher value-added models in this setting. We therefore had to address two key challenges. First, estimating teacher effects using observational data has to account for any sorting between students and teachers. Second, even if children are randomly assigned to teachers, the variance in test scores across students sums over the variances across classrooms, teachers, schools, and student-year shocks. Thus, we must decompose the variance over test scores into the variance of each of these components to properly estimate the variance of teacher quality. These two problems are exacerbated in our context because most children study in small schools with a single classroom per grade, and, in some cases, children from multiple grades may be taught in the same "multigrade" classroom. While this is a common scenario in Pakistan and other low-income countries,[1] it does not permit us to control for school-grade-year variation when we estimate TVAs or to estimate classroom-specific effects within a school-grade-year (a common first step to identifying the variance in school effects). We simply do not observe multiple teachers teaching the same grade in the same school in the same year, leading to legitimate concerns that existing methods may perform poorly in this setting.

In order to address the potential for sorting to teachers, we first show that controlling for a rich set of student characteristics, school-level student-teacher ratios, and whether the classroom is multigrade does not affect our estimates of TVA. The correlation between our baseline measures of teacher effectiveness with parsimonious controls and estimates using this augmented set of controls always exceeds 0.89. We additionally find that children learn as much in multigrade as in single-grade classrooms, which is of independent interest given the high prevalence of multigrade teaching in low-income countries.

Next, to assess sorting on unobserved characteristics, we examine children who switch schools during the four years of data collection in the spirit of Chetty, Friedman, and Rockoff (2014a). We show that the gain in test scores for a child who switches schools is predicted by the TVA of the teacher that she is matched to in the year of the switch,[2] suggesting that our estimated TVA indeed captures variation in teacher quality. We also show that the child's future teacher's TVA in the year of the switch does not predict her current teacher's TVA, suggesting that there is little systematic sorting that can lead to persistence in the quality of a child's teacher. This

[1] In the census of primary public schools for the Punjab province, average class sizes for grades 3, 4, and 5 in 2005 were 17, 16, and 13, and in our data, 13 percent of classrooms are multigrade, with 20 percent of teachers teaching a multigrade classroom at least once. More broadly, multigrade teaching is estimated to account for 26 percent of schools in Zambia, 36 percent in Burkina Faso, and 78 percent in India (Mulkeen and Higgins 2009, Blum and Diwan 2007).

[2] For this test, we do not include the children who switch schools in the estimation of the TVAs.

out-of-sample validation test is similar to Chetty, Friedman, and Rockoff (2014a), who exploit the entry and exit of teachers into a school and suggests that we can extend TVA estimation to smaller schools and classrooms.

Finally, our methods for decomposing variances in the test score distribution closely follow the literature. We estimate the variance of teacher effectiveness by estimating classroom-specific fixed effects, where one teacher is typically associated with multiple classrooms in the same school over time and use the covariances between these classroom effects to back out the variance of teacher and school effectiveness. Using these estimates, we then follow the procedure discussed by Kane and Staiger (2008) and produce empirical Bayesian estimates of TVA. These estimates are shrunk toward 0 (the mean of the distribution) to account for error in the TVA estimation, which would lead to attenuation bias when TVA is included on the right side of regressions.

Having shown that TVA estimation in this setting is methodologically sound, we then present three sets of findings. First, teachers in public schools have large and significant effects on children's learning. A 1 SD increase in TVA leads to a 0.15 SD increase in student test scores on average, with higher teacher effects for math and English compared to the vernacular, Urdu. Thus, moving a student from the fifth to the ninety-fifth percentile of the public sector TVA distribution increases test scores on average by 0.49 SD, which can be compared to an average annual test score gain of 0.33 SD for the cohort we observe across all four years of our sample.

Second, observed teacher characteristics explain no more than 5 percent of the within-district variation in TVA. Although measurement error implies that we can never fully explain variation in TVA with observed characteristics, the fraction that we can explain is strikingly small, especially since (i) our data include a rich set of teacher characteristics like teachers' test scores, and (ii) the variation in teacher qualifications in our data is considerably higher than in the United States. Nonetheless, we do find that the first two years of teaching experience are associated with a significant increase in TVA. Similar to recent research in sub-Saharan Africa, we also find that higher content knowledge is associated with higher TVA, a correlation that emerges clearly once we account for measurement error using multiple test scores from the same teacher (Bold et al. 2016 and Bietenbeck, Piopiunik, and Wiederhold 2018).

Third, public sector wages reward seniority and education—both of which have small associations with TVA—but not TVA directly. In contrast, rewards to seniority are one-third to one-fourth as high in the private sector. In striking contrast to the public sector, a 1 SD increase in TVA is associated with (at least) 49 percent higher wages in the private sector. Even in the absence of a formal testing regime, TVA is observable and can be rewarded, but the public sector does not have a mechanism to do so. Finally, the TVA of public sector teachers on temporary contracts, hired up to four years after a change in the hiring regime, was at least as high as that of permanent teachers, despite 35 percent lower salaries.[3] Both these findings point to a fundamental disconnect between TVA and wages in the public sector in our context.

[3] These findings are consistent with the fact that in 2003 teacher salaries in the public sector were five times higher than those in the private sector (Andrabi, Das, and Khwaja 2008), and by 2011 they were eight times as high, suggesting that teacher wages are not a binding constraint for hiring effective teachers in the public sector. This finding

The fact that TVA is only weakly observable when teachers are hired has led researchers to suggest a policy of selective tenure, where teachers are at first hired on a contract basis at lower wages before being confirmed as civil servants. Then, tenure is granted to the best performing after an observation period (Pritchett and Murgai 2006 and Muralidharan 2016). Our finding that contract teachers perform as well as permanent teachers, with substantial fiscal savings, suggests that such hiring policies with a contract period could be implemented in low-income countries. Following Staiger and Rockoff (2010), we use our data to simulate the impact of such policies, varying both the selectivity of tenure and the length of observation prior to the tenure decision. Like Staiger and Rockoff (2010), we show that such policies are most efficacious when they are highly selective, with only the observed top 20–40 percent of contract teachers receiving tenure after an observation period of 2 or more years. The steady-state yearly learning gains from these highly selective policies range from 0.05–0.07 SD, only slightly smaller than the maximal gains of 0.08 SD obtained by Staiger and Rockoff (2010) using US data. However, these steady-state gains can take a number of years to achieve. In our simulations, less than half the gains are realized in the first 15 years.

This paper contributes to the literature in three ways. First, our results add to the growing body of evidence showing the robustness of TVA estimates to different settings. Substantively, our results are similar to those found in the United States and Ecuador, another non-high-income country where researchers have been able to identify TVAs. This is true for the effects of moving a student from a low-quality to high-quality teacher (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Chetty, Friedman, and Rockoff 2014a; and Araujo et al. 2016), the low correlations between observed characteristics and TVA (Rivkin, Hanushek, and Kain 2005; Staiger and Rockoff 2010; and Araujo et al. 2016), and the positive effects of the first two years of teacher experience on students' outcomes (Rockoff 2004; Rivkin, Hanushek, and Kain 2005; and Araujo et al. 2016).[4] However, departing from the literature, we find a significant correlation between teacher content knowledge and TVA. We also find evidence of relatively low fade-out of teacher effects on student test scores when compared to Jacob, Lefgren, and Sims (2010), though our estimates are similar to those of Chetty, Friedman, and Rockoff (2014a).

Second, we are able to demonstrate that compensation in the public sector is uncorrelated with productivity. This mirrors the experimental evidence from Kenya and India (Duflo, Dupas, and Kremer 2011; Muralidharan and Sundararaman 2013; and Duflo, Dupas, and Kremer 2015), which shows that similar or greater levels of learning can be achieved by hiring contract teachers at lower wages as opposed to hiring tenured teachers. It also mirrors an audit study with Indian doctors (Das et al. 2016) that shows that compensation in the private sector is strongly linked to performance, while in the public sector, the correlation is zero. Similar to these papers, the extent of the disconnect between teacher pay and productivity that we identify in

complements research from Indonesia, where raising teachers' wages had no effect on their performance along the intensive margin (De Ree et al. 2018).

[4] Our results are also consistent with estimates of teacher effectiveness from Uganda that indicate that a 1 SD better teacher would increase student test scores by 0.14 SD (Buhl-Wiggers et al. 2017).

the public sector is remarkable. Relative to a public sector premia of 5 to 15 percent in OECD countries (Disney and Gosling 1998, Dustmann and van Soest 1998, and Lucifora and Meurs 2006), public sector wages are 5 times as large as private sector wages, and within the public sector, a decline in wages of (at least) 35 percent has no negative impact on productivity as measured by TVA.[5] Public sector wages are uncorrelated with TVA even though there is a strong correlation in the private sector. These large and significant misallocations in public sector pay in low-income countries could potentially allow for alternate recruitment policies that lead to greater learning at a lower cost.

Third, we simulate the impacts of an alternative recruitment policy of selective tenure (Pritchett and Murgai 2006, Staiger and Rockoff 2010, and Muralidharan 2016). These policies boost learning by 0.07 SD each year, with half the gains realized after 15 years.[6] Our estimates of the persistence of teacher effects suggest that an additional gain of 0.07 SD in each year will lead to a 0.17 SD gain at the end of 10 years of schooling. In addition to these average gains, these policies reduce the number of very poorly performing teachers, meaningfully improving the outcomes of students who would otherwise be taught by those teachers.

The remainder of our paper is organized as follows. Section I describes the setting and context, and Section II discusses the data. Section III discusses TVA estimation, the results of regressions of TVA on teacher characteristics, and the robustness of the TVA measures. Section IV discusses teacher hiring and compensation, including the link between teacher quality and teacher wages. Section V concludes.

I. Setting and Context

The data are from rural Punjab, Pakistan, the largest province in the country, with an estimated population of 110 million. The majority of parents in the province can choose between free public schools or paying for their child to attend a private school, and at the primary level, one-third of parents of enrolled children choose private schools.[7] Although funding for public schools has traditionally been small, in recent years, the government of Punjab has ratchetted up education budgets from 468 million dollars in 2001–2002 to 1.680 billion dollars in 2010–2011 (Ishtiaq 2012). Much of this expenditure is on recurring budget items, and similar

[5] Since we do not include future liabilities such as pensions in this accounting, the wage difference is a lower bound in our study.

[6] As Staiger and Rockoff (2010) discusses, these effects reflect the imperfect observability of TVA coupled with a discount for inexperience, both of which “drag down” average quality in the teaching workforce. If TVA is perfectly observable prior to the hiring decision, a policymaker could always choose to hire just the top 10 percent (say), leading to a pool of only the top 10 percent in the teaching workforce. But suppose it takes one year to observe productivity, and some permanent teachers retire each year. Then, in any given year, a fraction of the teaching workforce will be newly hired contract teachers whose expected TVA will be the mean TVA of the applicant distribution discounted for experience effects. Increasing selectivity improves the quality of the tenured pool but also increases the number of newly hired contract teachers. This trade-off establishes the optimal selectivity of the policy.

[7] Religious schools, or madrassas, account for 1–1.5 percent of primary enrollment shares, and their market share has remained constant over the last two decades (Andrabi et al. 2006).

to other low-income settings, teachers' salaries account for 80 percent of spending (UNESCO Islamabad 2013).[8]

In Pakistan, teachers are part of the civil service. They are recruited centrally (not locally), and once permanent teachers have been inducted into the workforce, there is virtually zero attrition. The civil service is considered a lifelong job (Chaudhury et al. 2006). Teachers can be transferred across schools, but both the recruitment and transfer process were not rule based at the time of our surveys. Once teachers are hired, salaries are determined through the Basic Pay Scale, which allocates teachers to "grades" based on their position (primary versus middle/secondary school teacher and regular teacher versus head teacher). Basic salaries that depend on grade and experience are supplemented with further allowances (see online Appendix Figure A1) (Idara-e-Taleem-o Aagahi 2013). One feature of teachers' wages that we emphasize is that these are not "lock-step" schedules with zero flexibility since teachers teaching at similar levels may receive different allowances, and there is some flexibility in the salary band.

Low attrition and poor accountability contribute to wide variation in observed teacher characteristics and behavior. In our data, 49 percent of public school teachers do not have a bachelor's degree, which partially reflects an increase in educational requirements over time (see online Appendix Figure A2). Consistent with low accountability, public school teachers *self-reported* being absent 3.2 days per month in 2003–2004, with days absent ranging from 0 at the tenth percentile to 6 at the ninetieth percentile.[9]

In sharp contrast to the public sector, teachers in the private sector are locally recruited and hired without the security of tenure. Their wages are determined through individual bargaining. Andrabi, Das, and Khwaja (2008) shows that teachers' wages in private schools were one-fifth of teacher salaries in public schools in 2003–2004, and public school salaries have only grown relative to private school salaries since then (online Appendix Figure A3).[10]

Like in most low-income countries, children's test scores are low, and by grade 3, children have barely mastered the grade K–1 curriculum. In our grade 3 sample, 84 percent can add 2-digit numbers, but only 69 percent can subtract double digits correctly. In English, only 2.4 percent can write a correct sentence with the word "play," and only 1 percent can write a sentence with the word "deep" (although a majority accurately match simple words to pictures). In Urdu, children stumble when reading paragraphs. From grade 3 until the end of primary school (grade 5), there is considerable variation in how much children learn. On average, children in public schools gained 0.78 SD, but the bottom tenth percentile of learning gains was only 0.09 SD, while the top tenth percentile of learning gains was 1.41 SD.

[8] In a study of 55 low-income countries, Bruns, Mingat, and Rakotomalala (2003) shows that teacher salaries account for 74 percent of recurring spending by the government on education.

[9] For comparison, the absence rate of an average teacher in the United States is 5 percent and is only 3.5 percentage points higher for schools whose proportion of African American students (a marker associated with disadvantage in the United States) is in the ninetieth percentile (Miller 2012).

[10] Similar wage gaps have been documented in Colombia, the Dominican Republic, the Philippines, Tanzania, Thailand, and India (see Jimenez, Lockheed, and Paqueo 1991 and Muralidharan and Kremer 2008).

II. Data

We use data collected across four rounds (2003 to 2007) of the LEAPS survey. The 2003 sample included 496 public schools in 112 villages of three districts in the province of Punjab,[11] with an additional 23 primary public schools entering the sample over the next four years. For our purposes, two parts of the data collection are key. First, a teacher roster was completed for all teachers within the school in each year of the survey. This roster included sociodemographic data on teachers (gender, age, educational attainment) and in the fourth round, month-level data on when the teacher began teaching. We use variables from the teacher roster to look at the differences between teachers in demographic characteristics, salaries, and subject knowledge. Online Appendix Table A1 provides summary statistics for these characteristics across the four rounds of the survey for both public and private school teachers. Online Appendix Table A2 reports additional summary statistics for the schools in the sample, including their size, number of teachers, and facilities.

Most schools in our sample have one or fewer teachers per grade, and we observe multiple teachers teaching in a grade in 26 percent of schools. In online Appendix Table A3, we report the number of public sector teachers observed teaching each combination of grades 3, 4, and 5. Because teachers teach multiple grades, it is possible for a teacher to be observed teaching two or more grades even if they are observed once. For example, 8 teachers are only observed once but teach grades 3, 4, and 5 simultaneously. We report the teacher counts for both the full sample and a restricted sample that excludes teachers who ever appear to teach the same class in concurrent years (that is, more than 25 percent of their students in year t also were taught by them in year $t - 1$). While many teachers are only observed once, a large number of teachers are observed two or more times, even when teachers who teach essentially the same students in subsequent years are excluded (195 total teachers observed teaching grades 3, 4, or 5).

In the second part of our data collection, to assess learning outcomes, LEAPS tested children in the survey schools. English, Urdu, and mathematics tests were administered to children in grades 3 through 6 between 2004 and 2007.[12] In the first year of data collection, only classrooms with grade 3 students were tested. In subsequent years, those children were followed to new classrooms, and an additional cohort of third graders was added in year three and followed in year four. Online Appendix A discusses the implementation and scoring of the tests. Here, we note that (i) the tests were low stakes and designed by researchers to maximize precision over a range of abilities in each grade, and (ii) scores could be equated across years using a set of linked questions in each year together with item response theory

[11] The three districts were chosen on the basis of an accepted stratification of the province into the better-performing north and central regions and the poorly performing south. The project was designed as part of a study of the rise of private schooling, and as a result, all the villages included in the study were randomly selected from a list frame of villages with at least one private school when the study began in 2003. As these villages tend to be larger and wealthier, the sample is representative of 60 percent of the rural population in the province of Punjab (Andrabi et al. 2008).

[12] While we include grade 6 test scores in our analysis, we mainly focus on primary school teachers and grades 3 through 5 in this paper because only the oldest cohort of students attends grade 6 and only in 2007. Therefore, we do not have multiple years of observation for most grade 6 teachers.

(IRT), as in Das and Zajonc (2010). These test equating methods allow us to score all children in all years on the same knowledge scale (whose distribution is close to a normal distribution with mean 0 and standard deviation 1) in a comparable fashion and therefore compare TVA estimates to annual learning gains in the sample.[13] Online Appendix Table A1 documents test score gains by year over the four rounds of testing in the panels of public school students.

On the day that children were tested, we also asked teachers to take the same test as the children.[14] Online Appendix Figures A4 and A5 show the histograms of teacher test scores in public and private schools, and although there are ceiling effects, particularly in math, where 14 percent of public teachers achieve the maximum, there is also a great deal of variation. We verify that the teachers' test scores can be equated to child test scores by predicting item characteristic curves based only on estimation of item parameters from the child test sample and assessing whether teacher responses fit item response curves for each test item. Online Appendix A shows that this equating process works well. There are only two to three items in each test for which teacher responses do not correspond to the characteristic curve estimated with child scores.

Equated to the child test score distribution, the mean public teacher's average test score is 3.04 SD higher, but the fifth percentile of teachers is only 1.91 SD higher than the average tested child. Since the process of testing teachers was repeated each year, for teachers who we observe multiple times, we use the multiple observations of teachers' test scores to correct for measurement error.[15] Online Appendix Table A4 shows that teacher characteristics at most explain 9 percent of the variation in content knowledge for public school teachers, and reassuringly, having a bachelor's degree is robustly correlated with content knowledge.[16] However, the correlation is relatively small (0.2 to 0.3 SD), which could either reflect the quality of the degree or "learning on the job" among those without a degree.

Teacher quality is identified in two different ways following the TVA literature (Rockoff 2004; Kane and Staiger 2008; and Chetty, Friedman, and Rockoff 2014a). When we generate estimates of teacher quality that are outcomes in regression equations, we regress student test scores on a function of their lagged test scores, survey round, grade, and teacher fixed effects. Teacher value added is then the estimated teacher fixed effect. The panel structure of the data, where both students and teachers are observed multiple times, is important for identification. To be included in the value-added calculations, students must be observed at least

[13] The variance of the test score distribution varies only marginally across years in our sample. Therefore, the TVA estimates can also be meaningfully compared to the cross-sectional variation in any year, which is approximately 1 SD.

[14] Since the test administered to teachers was the same as the test administered to students, unlike other measures of teacher knowledge such as the commonly used Praxis test, this measure is ideal for assessing teachers' command of the content knowledge required for the classrooms in which they were teaching.

[15] Teachers were only tested if they taught in the grade that the students should have been enrolled in if they were progressing through the school system on track. So, in year one, grade 3 teachers were tested; in year two, grade 4 teachers were tested; in year three, grade 3 and grade 5 teachers were tested; and in year four, grade 4 and grade 6 teachers were tested. Thus, some teachers were tested multiple times if they taught grades of interest in multiple years. Only 4 percent of teachers who should have been tested do not have at least one test score.

[16] The regressions for private school teachers yield qualitatively similar results.

TABLE 1—SOURCES OF VARIATION IN TEACHER VALUE-ADDED CALCULATIONS FOR THE PUBLIC SECTOR

	Number of teachers (1)	Number of students (2)	Teachers in schools with >1 teacher with tested students (3)	Students in schools with >1 teacher with tested students (4)
Round 1	486	8,340	4	131
Round 2	593	9,327	214	3,290
Round 3	1,007	16,946	884	15,320
Round 4	1,103	15,357	812	12,610

Notes: This table presents the breakdown of the data used to calculate within- and across-school TVAs in the public sector. Within-school TVAs require teachers to teach in schools where more than one teacher has tested students (so that the mean school effect is not equal to the sole teacher's TVA). The sample of students driving variation in the within-school TVAs are the students who attend schools where more than one teacher has tested students.

twice across consecutive years since they require lagged test scores to control for selection. To separate correlation in student outcomes within years from TVA, at least some teachers must also be observed across years so that round fixed effects are identified. When we generate estimates of teacher quality that we include as explanatory variables in regressions, we follow the empirical Bayes method proposed by Kane and Staiger (2008) and shrink our estimates toward zero to account for measurement error. As before, the panel structure of the data is essential to account for lagged student test scores and grade- and year-level shocks to test scores.

We observe a total of 1,756 public school teachers in at least one round of the data linked to 22,857 unique public school students. Online Appendix Figure A6 documents this variation, and Table 1 provides more information on the sources of variation for the TVA calculations. Several limitations lead to a smaller effective sample size. We estimate TVAs for 1,533 teachers since the TVA estimation does not include students if they were not observed in the prior year. In particular, this means that we cannot calculate TVAs for teachers who were only observed in the first round of the data. When we correlate TVAs with teacher characteristics, our sample is further reduced to 1,383 teachers because detailed data on when a teacher started teaching, which allows us to include our experience controls, was only collected in the fourth round of the LEAPS study.

To account for unobservable variables, we can also de-mean TVA estimates at the school level. Since we do not observe teachers in more than one school, we cannot separately identify pure school effects as opposed to a school simply having better teachers on average. Therefore, the de-meaned TVAs should be interpreted as a within-school ranking of teacher quality, and if teachers' marginal products are equated within schools or teachers sort into schools, the de-meaning will absorb part of the true teacher value-added. Demeaning at the school level requires that more than one teacher was observed in the school over the course of the study, and columns 3 and 4 of Table 1 document this variation. TVAs for teachers in the 158 public schools where only one teacher was ever observed with tested students are left out of the within-school TVA sample; these teachers account for 2,357 child-year observations or 1,771 unique children.

III. Teacher Value Added

We divide our discussion of TVA into four sections. We first establish the importance of teachers in our context by estimating the variance of teacher effects. This allows us to compute the gains (for instance) of moving a student from a teacher at the fifth percentile of TVA to the ninety-fifth percentile of TVA. We discuss the methods and results from this exercise in Section IIIA. We then turn to the direct estimation of TVA and its correlates in Sections IIIB and IIIC, followed by a series of robustness tests and extensions in Section IIID.

A. *Teacher Effectiveness: Estimating the Variance of Teacher, School, and Classroom Effects*

Estimation.—To measure the variance of teacher, school, and classroom effects in our data, which are indicative of how much test scores would increase if a student were moved to a 1 SD better teacher, school, or classroom, we follow Araujo et al. (2016). We first assume

$$y_{isjgt} = \sum_a \beta_a y_{i,t-1} I_{it}(grade = a) + \theta_{jst} + \theta_s + \theta_{js} + \alpha_t + \mu_g + \upsilon_{isjgt},$$

where i denotes a student, t denotes a testing year, s denotes a school, g denotes a grade, and j denotes a teacher. Variables y_{isjgt} are then test scores in math, Urdu, and English; $I_{it}(grade = a)$ is an indicator variable equal to 1 if a student i attends grade a in year t; β_a is the grade-specific effect of lagged test scores, which allows past ability and investments to affect current performance; α_t is a year-specific shock for year t; μ_g is a grade-specific shock for grade g; θ_{jst} is a classroom-specific effect for a teacher j in school s in year t; θ_{js} is a time-invariant teacher-specific effect; θ_s is a time-invariant school-specific effect; and υ_{isjgt} is a student-year-specific idiosyncratic shock. Our objects of interest are σ^2_{jst}, σ^2_{js}, σ^2_s, and σ^2_υ, which are the variances of the classroom, teacher, school, and individual idiosyncratic shocks, respectively.

To measure the variance of these effects, we first generate estimates of classroom effects by estimating the regression

$$y_{isjgt} = \sum_a \beta_a y_{i,t-1} I_{it}(grade = a) + \delta_{jst} + \alpha_t + \mu_g + \upsilon_{isjgt}, \tag{1}$$

where δ_{jst} is the classroom-level fixed effect (equivalent to a teacher-year fixed effect in our sample), which subsumes the classroom-level shocks such as having a more disruptive student in the classroom in a given year, the teacher effect, and the school effect. Then, $\sigma^2_s = \text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{j'st})$, the covariance between the classroom effects for two classrooms in the same school taught by teacher j and teacher j'. To estimate the variance of the teacher effects, we exploit the fact that $\sigma^2_s + \sigma^2_{js} = \text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{jst'})$ (McCaffrey et al. 2009). Having already estimated the variance of the school effects, we can then subtract $\text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{j'st})$ from $\text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{jst'})$ to identify the variance of teacher effects. Finally, to identify the variance of classroom effects, we estimate $\text{var}(\hat{\delta}_{jst}) = \sigma^2_{jst} + \sigma^2_{js} + \sigma^2_s + \phi$, where ϕ is the sampling bias that comes from the fact that we observe estimates of δ_{jst} instead of the true values. In online

Table 2—Effect of a 1 SD Improvement in School, Teacher, Classroom, and Individual Effects on Student Test Scores in the Public Sector

	Math (1)	Urdu (2)	English (3)	Average (4)
Panel A. Full sample				
Classroom	0.27	0.31	0.25	0.28
School	0.15	0.14	0.22	0.17
Teacher	0.21	0.06	0.17	0.15
Individual	0.52	0.50	0.48	0.50
Panel B. Restricted sample				
Classroom	0.30	0.35	0.32	0.32
School	0.13	0.06	0.19	0.13
Teacher	0.19	0.11	0.15	0.15
Individual	0.52	0.50	0.49	0.50

Notes: This table reports the effect of receiving a 1 SD higher classroom, school, teacher, or individual idiosyncratic shock on students' subject-level test scores, as well as the average effect across the three. Test scores are estimated with IRT and measured in standard deviations. To arrive at these numbers, we use equation (1) to estimate teacher-year fixed effects in the panel dataset of student test scores. Denote $\hat{\delta}_{jst}$ the teacher-year fixed effect for teacher j in school s in year t. Then, the school variance is $\text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{j'st})$, the teacher variance is $\text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{jst'}) - \text{cov}(\hat{\delta}_{jst}, \hat{\delta}_{j'st})$, the classroom variance is the variance of $\hat{\delta}_{j'st}$ minus the sampling bias, which we solve for analytically in online Appendix B, and the individual variance is the variance of the residuals. In panel A, the sample includes all students and teachers in public schools. In panel B, the restricted sample excludes teachers who are ever observed teaching two classes of students who appear to be the same in two consecutive years (25 percent or more of the students in year t were taught by the same teacher in year $t-1$).

Appendix B, we derive an expression for ϕ, which can be estimated with our data. This gives us enough information to identify σ^2_{jst}. Finally, we estimate σ^2_v as the variance of the residuals.

One limitation of this strategy is that we cannot allow school shocks to vary across years since, unlike Araujo et al. (2016), we often do not observe two teachers teaching in the same school within the same year. Therefore, our estimate of σ^2_s generally relies on taking the covariance of the classroom effect estimates for two different teachers in school s in two different years. This is a necessity in many low-income countries, like Pakistan, where there is often only 1 teacher and 1 class per grade in the modal school (see online Appendix Figure A7).

Results.—Columns 1–3 of Table 2 report the effect of a 1 SD improvement in a student's classroom, school, or teacher on a student's test scores in math, Urdu, and English. The final column reports the average across these three effects. In panel A, we use the full sample of students. Holding other components fixed, a 1 SD better classroom leads to 0.28 SD (student-level) higher performance, while a 1 standard deviation better school leads to 0.17 SD higher test scores. Teacher effects are almost as large: a 1 SD better teacher leads to 0.15 SD higher test scores, and moving a student from a teacher in the fifth to the ninety-fifth percentile of the TVA distribution would lead to a 0.49 SD increase in mean test scores.[17] One concern is that

[17] To arrive at this number, note that a teacher who is at the fifth percentile is 3.3 SDs worse than one at the ninety-fifth percentile. Therefore, the total test score effect is $3.3 \times 0.148 = 0.49$, where 0.148 is the three-decimal-point effect of a 1 SD better teacher.

the classroom effects in our data are not truly independent if a teacher remains with the same students across grades, upwardly biasing estimates of the teacher effects. Therefore, in panel B, we re-estimate these effects restricting the sample to teachers whose current class has less than 25 percent of the same students as their class in the previous year. Our estimates for this subsample are quite similar, and the estimates of the average teacher effects are virtually identical to those in the previous panel.[18]

We can also use these estimates to compute the percentage of the variance in test scores that is explained by the variance in teacher effects. According to our estimates in panel A of Table 2, the teacher effect accounts for 8 percent of the total variance in test scores due to schools, teachers, classrooms, and students in English, 11 percent in math, and 1 percent in Urdu.[19] This compares to 3 percent in English and 7 percent in math in Chetty, Friedman, and Rockoff (2014a). This again suggests that the portion of the variance in test scores due to teacher effects is similar but perhaps somewhat higher in Pakistan compared to the United States. Like Chetty, Friedman, and Rockoff (2014a), we also find that teachers account for less of the variance in the vernacular, which students may learn at home (in our case, Urdu), and more of the variance in subjects like English and mathematics, where teaching at school and at home are less likely to be substitutes. Overall, our estimates of the variance of the classroom and teacher effects are larger than the estimates in Ecuador (Araujo et al. 2016), similar to those in Uganda (Buhl-Wiggers et al. 2017), and on the high end of estimates in the United States (see Rockoff 2004 and Chetty, Friedman, and Rockoff 2014a).

B. *Estimating Teacher Value Added*

In our analysis of teacher productivity, we include TVA estimates (which are the predicted test score gains for a student from being assigned to a given teacher) on both the right and left side of regression equations. Since measurement error in TVA estimates leads to attenuation bias when TVA is included on the right side of the regression, we estimate TVA in two different ways. For cases where we include TVA on the left side of the equation, following Rockoff (2004), we estimate TVA as a teacher fixed effect. This method is similar to the methods of Harris and Sass (2006); Kane and Staiger (2008); Chetty, Friedman, and Rockoff (2014a); and Chetty, Friedman, and Rockoff (2014b), but unlike some of these approaches, the fixed effect approach allows us to estimate TVA even for teachers who are only observed once in the data, as long as the children have (at least) two test scores in consecutive years. To compute TVA, we estimate the following regression on our full set of public school teachers including all child-year test score observations:

$$y_{ijgt} = \beta_0 + \sum_a \beta_a y_{i,t-1} I(grade = a) + \gamma_j + \alpha_t + \mu_g + \epsilon_{ijgt}, \tag{2}$$

[18] Our methodology here follows the literature, which assumes either that teacher quality is time-invariant or that it is stationary. However, teacher quality may be affected by experience, which would violate these assumptions. To verify that this does not qualitatively affect our estimates, we re-estimate the baseline classroom fixed effect regressions including fixed effects for years of teacher experience. This has little effect on our estimate of the average effect of a 1 SD better teacher, which declines from 0.148 to 0.142.

[19] The total variance for each subject is simply the sum of the variances for the teacher, classroom, school, and individual components, and these variances are the squares of the standard deviations reported in Table 2.

where y_{ijgt} is again a student's test score, γ_j is the fixed effect for teacher j, α_t is the fixed effect for round t, and μ_g is the fixed effect for grade g. Then, γ_j is the TVA, equivalent to the underlying unexplained variance in test score gains associated with students having the same teacher. Since student test scores are standardized using IRT, TVA measures are in student test score standard deviations. As is conventional in the TVA literature, we control for year-specific and grade-specific shocks, as well as lagged test scores, which are allowed to affect students in different grades differently. These account for students' prior human capital attainment and the selection of students to teachers. Since we do not observe teachers changing schools, these initial fixed effects estimates include both teacher and school effects.[20] Here, our key identifying assumption is that, conditional on the control variables (including the rich function of lagged test scores), students who learn more or less quickly don't selectively sort to different teachers.[21]

For cases where TVA will be included on the right side of the regression, we follow the procedure outlined in Kane and Staiger (2008), which computes empirical Bayesian estimates of TVA. Using our average variance estimates[22] from panel A of Table 2 and our estimates of the classroom fixed effects in equation (1), we form weighted averages of the classroom effects for each teacher, weighting the classroom fixed effects by their inverse variance. We then multiply these weighted averages by a shrinkage factor that is equal to the ratio of the signal variance to the total variance.[23] This shrinkage factor effectively shrinks the estimates toward zero, counteracting attenuation bias from measurement error.

Additionally, while we cannot separately identify school effects, we also construct within-school estimates of TVA to remove any bias that the lagged test scores fail to control for from students sorting into schools. In the case of our fixed effects approach, to construct these estimates, we de-mean the teacher fixed effects at the school level. In the case of the empirical Bayesian approach, we first de-mean the

[20] Our TVA estimates do not capture teachers' heterogeneous effects on different students. In reality, such heterogeneity may be important for students' outcomes. See, for instance, Aucejo (2011) and Bau (2019) on match-specific school and teacher quality and Dee (2007); Hoffmann and Oreopoulos (2009); Antecol, Eren, and Ozbeklik (2015); and Muralidharan and Sheth (2016) on heterogeneous effects by gender. If teachers have heterogeneous effects on students, our TVA measures can be thought of as capturing the average effect across students.

[21] Like Kane and Staiger (2008) and Chetty, Friedman, and Rockoff (2014a), we do not include child fixed effects to account for additional unobservable selection of students to teachers. Online Appendix C discusses issues that could arise with the inclusion of child fixed effects in a context where children switch teachers infrequently. In that case, identification with child fixed effects would be based on a smaller subsample of students who are observed with multiple teachers over time. Then, measurement error in how teacher codes are entered into the data will lead to false switchers—students who appear to be switching teachers but actually are not. Even with a small number of false switchers, this could lead to substantial biases in the estimation of TVA by inducing spurious correlations between the TVA of teachers with similar ID numbers.

[22] We use the average variance estimates instead of the subject-specific estimates because the average estimates across subjects are more stable across estimating samples. For example, the variance estimate for teachers is 0.022 in both the restricted and full sample, but it is 0.004 in the full sample for Urdu and 0.011 in the restricted sample.

[23] Formally, the weighted average of the classroom effects for a teacher j is given by $v_j = \sum_i w_{jt} \hat{\delta}_{jt}$, where $w_{jt} = h_{jt}/\sum_t h_{jt}$, $h_{jt} = 1/(\hat{\sigma}^2_{jst} + \hat{\sigma}^2_{\upsilon}/n_{jt})$, and n_{jt} is the number of students taught by teacher j in year t. Then, the shrinkage factor is given by $\left(\hat{\sigma}^2_j + \hat{\sigma}^2_s\right)/\left(\hat{\sigma}^2_j + \hat{\sigma}^2_s + \left(\sum_t h_{jt}\right)^{-1}\right)$.

classroom effects used to construct the TVA estimates at the school level and then adjust the shrinkage factor accordingly.[24]

Implementing the procedures described in this section results in TVA estimates for 1,533 teachers. In the next sections, we first examine TVA's correlations with teacher characteristics. We then use the TVA estimates to assess the link between productivity and wages in both the public and private sectors, using the same procedures to estimate TVA in the private sector.

C. *Teacher Value-Added Results*

Teacher Characteristics and TVA.—Using our TVA estimates, we estimate the association between TVA and observed teacher characteristics for public school teachers using the following specification:

$$TVA_j = \beta_0 + \Gamma X_j + \alpha_d + \varepsilon_j,$$

where TVA_j is a teacher j's average value-added over math, Urdu, and English; X_j consists of teacher characteristics, including an indicator variable for some training, an indicator variable for having a bachelor's degree or greater, an indicator variable for having three or more years of experience in 2007, an indicator variable for female, an indicator variable for whether a teacher is local, an indicator variable for whether a teacher has a temporary contract, controls for age and age squared, and in some specifications, controls for a teacher content knowledge; and α_d is a fixed effect for district d. In some specifications, we also include a school fixed effect. Since TVA is on the left-hand side of this expression, we use our fixed effects measures for the outcome variable instead of the empirical Bayes measures.

Table 3 presents the results from this specification. Column 1 reports the means of the covariates of interest. Columns 2 and 3 report regression results without controlling for teachers' own test scores, and Columns 4–7 include different measures of teachers' test scores with the goal of reducing measurement error in teacher content knowledge.[25] Like in the United States, we are never able to explain more than 5 percent of the variation in mean TVA.[26] We find no significant, positive correlation between TVA and education (measured as whether a teacher has a bachelor's degree) and whether the teacher has some training. Nevertheless, two correlations are statistically significant and of particular interest.

[24] The new shrinkage factor is $\hat{\sigma}_j^2/\left(\hat{\sigma}_j^2 + \left(\sum_t h_{jt}\right)^{-1}\right)$ since the variance of the school effect is no longer part of the signal.

[25] In these regressions, we cannot identify the association between teacher gender and TVA when we include school fixed effects, and, therefore, we do not report an association between gender and TVA for these specifications. This is because public schools in Pakistan at the time of the data collection were not coeducational, and less than 5 percent (29) had both a male and female teacher during the sample period.

[26] We arrive at 5 percent by first regressing mean TVA on district or school fixed effects, then regressing mean TVA on the fixed effects and teacher characteristics, and then calculating the difference in the adjusted R^2s.

Table 3—Relationship between Teacher Characteristics and Mean Teacher Value Added for Public School Teachers

	Covariate mean (1)	Mean TVA (2)	Mean TVA (3)	Mean TVA (4)	Mean TVA (5)	Mean TVA (6)	Mean TVA (7)	Mean test score (8)
Female	0.449	0.068	N.A.	0.079	N.A.	0.085	N.A.	
		(0.026)		(0.026)		(0.032)		
Local	0.273	0.025	0.014	0.023	−0.001	0.017	−0.111	
		(0.025)	(0.030)	(0.028)	(0.049)	(0.036)	(0.079)	
Some teacher training	0.904	0.027	−0.048	−0.096	−0.214	−0.125	−0.598	
		(0.056)	(0.074)	(0.076)	(0.125)	(0.118)	(0.329)	
Has bachelor's degree or better	0.514	0.024	0.009	0.007	0.005	−0.069	−0.175	
		(0.025)	(0.031)	(0.033)	(0.058)	(0.046)	(0.108)	
Had >3 years of experience in 2007	0.868	0.076	0.085	0.037	0.158	−0.047	0.180	
		(0.038)	(0.051)	(0.047)	(0.096)	(0.072)	(0.274)	
Temporary contract	0.229	0.017	0.059	−0.017	0.058	0.020	0.067	
		(0.036)	(0.048)	(0.043)	(0.082)	(0.057)	(0.105)	
Mean teacher knowledge	3.041			0.088	0.018	0.294	0.330	
				(0.024)	(0.038)	(0.073)	(0.176)	
More than 2 years experience	0.925							0.079
								(0.029)
Fixed effects		District	School	District	School	District	School	NA
Endogeneity test statistic						15.320	4.425	
F-statistic		1.978	0.624	3.072	0.698	73.428	2.741	
Adjusted R^2		0.224	0.454	0.228	0.417	0.070	0.039	0.633
Observations		1,383	1,383	919	919	622	622	26,472
Clusters		471	471	469	469	440	440	569

Notes: This table reports estimates of the association between TVA and teacher characteristics. The association between female and TVA in the public sector cannot be credibly estimated in the presence of school fixed effects because the public sector is not coeducational. Very few public schools (29) are observed with both male and female teachers over the course of the sample. The first column reports the means of the covariates of interest at the teacher level, and the remaining columns are regressions whose header specifies the dependent variable. For columns 2–7, observations are at the teacher level and characteristics are time invariant. In these columns, we regress estimates of TVA for mean student test scores on time-invariant teacher characters and district (even columns) or school (odd columns) fixed effects. In columns 4–5, mean teacher knowledge is the average of a teacher's test scores on the same tests taken by students across all the tests teachers took. In columns 6–7, we instrument for the teacher's mean score in the first tested year with the mean score in the second year. The sample size is reduced from column 3 to column 4 and from column 5 to column 6 because not all teachers were tested and not all teachers were tested in multiple years. Column 8 identifies the association between teacher experience and student test scores, controlling for lagged test scores interacted with the child's grade, year fixed effects, and grade fixed effects in line with the controls in equation 1. Observations for this column are at the student-year level. The F-statistic is for an F-test of all the covariates in columns 2–5. In columns 6 and 7, it is the F-statistic from the first stage of the instrumental variables regression. Standard errors are clustered at the school level.

First, content knowledge, measured as a teacher's average test scores over all subjects across all years, is significantly correlated with the estimated mean TVA (column 4). Columns 6 and 7 further reduce measurement error by instrumenting for the teacher's first cross-subject average test score with her second.[27] Our preferred IV specification suggests that a 1 SD increase in teacher test scores increases TVA by

[27] The inclusion of teacher test scores causes the sample size to fall since not all teachers took the test. Instrumenting for teachers' first test scores with their second test scores further reduces the sample since even fewer teachers took the test twice. To demonstrate that the large and significant effects of content knowledge documented in Table 3 are not driven by the selection of a sample of teachers who were tested at least once (columns 4 and 5) or twice (columns 6 and 7), in online Appendix Table A5, we also re-estimate our baseline specifications (columns 2

0.29–0.33 SD, which is higher than the effects estimated by Metzler and Woessmann (2012) (0.1 SD in Peru) and the effect of IQ (0.04 SD) estimated by Araujo et al. (2016).[28] However, it is similar in magnitude to the effects estimated by Bold et al. (2016) (0.23 SD–0.54 SD in Africa) and Bietenbeck, Piopiunik, and Wiederhold (2018). We speculate that this positive relationship is driven in part by the large variation in teacher knowledge levels in Pakistan, where a standard deviation in the teacher test score distribution in math is nearly as large as a standard deviation in the student test score distribution (0.87 versus 1), and a teacher at the fifth percentile scores below a student at the ninety-fifth percentile in math on a test designed for primary school students.

Second, experience in the first 0–2 years of teaching is also associated with higher TVA. Columns 2–7 treat each teacher as a single observation, but unlike the other characteristics in these regressions, experience is not time invariant. Since we observe teachers multiple times at different levels of experience, we can use our panel data to better identify experience effects for public school teachers. In column 8, an observation is a student-year, and we re-estimate equation (2) but now replace the teacher fixed effect with an indicator variable for having greater than two years of experience. The results in column 8 suggest that the outcomes of students of teachers with more than two years of experience are 0.08 SD better than those of other students.[29]

D. *Teacher Value Added: Robustness and Extensions*

In this subsection, we report the following robustness tests of the validity of the TVA estimates and extensions. First, if lagged test scores are insufficient to account for selection, our TVA estimates are biased. We assess this bias and present a test in the spirit of Chetty, Friedman, and Rockoff (2014a) suggesting our estimates are unbiased. Second, our estimates may be unstable due to sampling variability. In the absence of administrative data, TVA estimates in low-income countries will likely be based on small sample sizes. We therefore assess how our TVA estimates vary across subsamples of the data. Third, we assess whether multigrade teaching affects our TVA estimates. Finally, a high VA teacher affects learning both contemporaneously and in the future. In the United States, Jacob, Lefgren, and Sims (2010) shows that teacher effects fade out fairly rapidly. We therefore assess the degree of fade-out in the context of our data.

---

and 3) in these subsamples. As before, there are no statistically significant correlations between the other teacher characteristics and mean TVA.

[28] Teacher knowledge is significantly correlated with TVA, but having a bachelor's degree is not. Online Appendix Table A4 shows that having a bachelor's degree is only associated with a 0.22 SD (see column 8) increase in a teacher's average test score. Based on the IV results, the implied effect of a bachelor's degree would be 0.06 SD, which is not significantly different from the estimated effect of 0.024 (SE = 0.025) (see column 2 of Table 3).

[29] We note that Wiswall (2013) finds evidence that experience effects reemerge in the United States in later years. Therefore, following Wiswall (2013), we also semi-parametrically estimate marginal experience effects for each additional three years of teaching for math, Urdu, English, and mean test scores. Online Appendix Figure A8 graphs these estimates with an omitted category of more than zero years of experience. The findings in this graph again show that the first two years of experience matter non-linearly, and that, while the same pattern occurs for all three subjects, it is particularly strong for mathematics.

Assessing Bias Using Observables.—We first assess the extent of the bias by checking whether a rich set of additional controls for student and school characteristics in our TVA estimation alters our estimates.[30] Even though the new TVAs use substantially more information about students' socioeconomic status and school-level inputs (the controls jointly explain 11 percent of the variation in mean test scores), the correlations with our original, cross-school TVA estimates are 0.95 for English, 0.91 for math, and 0.91 for Urdu (see online Appendix Table A6). In columns 3 and 4, we also implement the methodology proposed by Altonji and Mansfield (2018), which shows that classroom-level means of observable student characteristics can proxy for unobservable characteristics related to student outcomes. Thus, we also include controls for classroom-level mean household assets and mean lagged test scores. With these controls, the covariates now explain 43 percent of the variation in test scores. Despite the fact that Altonji and Mansfield (2018) shows that these estimates are lower-bound estimates of the teacher effects, the correlations between the revised and original TVA estimates are now 0.70–0.77 for the across-school estimates and 0.69–0.79 for the within-school estimates.

Assessing Bias Using Switchers.—Our next test is an out-of-sample prediction test following Chetty, Friedman, and Rockoff (2014a). We focus on children currently enrolled in a government school who switched schools and assess whether test scores for the student can be fully predicted by the TVA of the new teacher (controlling for lagged test scores). The specific regression specification is

$$(3)\quad \textit{test score}_{ijt} \;=\; \beta_0 + \beta_1 TVA_j + \sum_a \beta_a y_{i,t-1} I\big(grade \;=\; a\big) + \alpha_t + \alpha_s + \mu_g + \varepsilon_{ijt},$$

where $\textit{test score}_{ijt}$ is the test score of a student i with a teacher j in year t, which can be in math, Urdu, or English or the average across all three; TVA_j is the value added of a student's teacher in the relevant subject; and α_s, α_t, and μ_g are school, year, and grade fixed effects, respectively. The sample consists of students who are in a new school in period t. Because we limit the sample to school switchers, β_1 will not be influenced by common shocks at the school level that are correlated over time, and if TVA is unbiased, in expectation, our estimate of β_1 should be equal to 1.

Our test proceeds in two steps. We first ensure that β_1 is not biased by selection between students and teachers. If students systematically sort to teachers, even when they switch schools, and the controls in the TVA estimation do not adequately control for this sorting, the TVA of the switcher's teacher after they switch schools should be correlated with the TVA of their current teacher. To assess whether this is the case, we regress the switchers' teachers' TVAs in the year prior to the switch on the TVA of their teachers after the switch.[31] Figure 1 plots the event study version of

[30] Our alternate estimates control for age; gender; the student's household assets index; parental education; two indices of school facilities that vary over time; time-varying, school-level student-teacher ratios; and whether the child was in a multigrade classroom (defined as two or more grades being in the classroom with at least 5 percent of students from each of the grades).

[31] Focusing on children who switched schools ensures that our test will not find spurious correlations between future and current TVAs due to the fact that school-grade level shocks to the current teachers' students' outcomes will effect the lagged test scores used to calculate the future teachers' TVAs, as described by Chetty, Friedman, and Rockoff (2015).

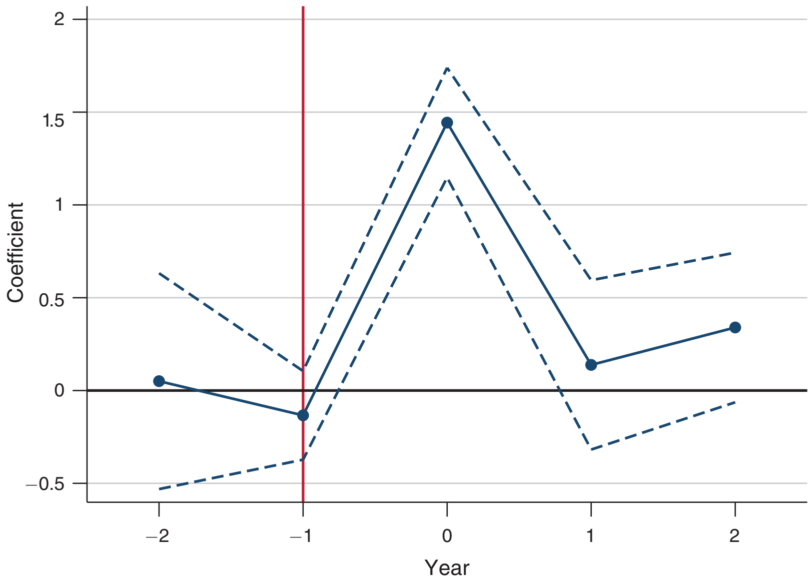

FIGURE 1. EFFECT OF MEAN TVA AT $t = 0$ FOR SCHOOL SWITCHERS ON MEAN TVAs IN OTHER YEARS

Notes: This figure plots the event study analog of Table 4, which estimates the effect of a student's future teacher's TVA after a school switch on her current teacher's TVA. The sample consists of public school students who switch schools. $t = 0$ is normalized to be the year a switch occurs, and the figure plots the coefficient of the empirical Bayes measure of mean TVA at $t = 0$ in a regression of the mean TVA of a student's teacher in year t on her teacher's TVA in year 0. The regressions control for fixed effects for the school a student attends in $t = 0$.

this regression. Here, we use a sample of students who switch schools and attend a public school in the year that the switch occurred and drop students who are taught by the same teacher multiple times. We regress the mean TVA of a student's teacher in year t, where t is normalized so that the school switch takes place at $t = 0$ (for $t = -2, \ldots, 2$), on the shrunk mean TVA of her teacher at time $t = 0$, allowing for different coefficients by t and controlling for school fixed effects using the school students attended at $t = 0$. The coefficient for this relationship at $t = 0$ should be mechanically greater than 1 because the TVA on the right side of the regression has been shrunk. For the remaining years, the coefficients are indistinguishable from 0, consistent with the hypothesis that students are not sorting systematically to teachers. In other words, the fact that a student matches to a high/low TVA teacher at the time of the switch has no predictive power for the TVA of the other teachers she was matched to either prior to or after the switch. Table 4 reports the regression results from regressing a student's current teacher's TVA on her future teacher's TVA for the year before the switch took place and shows that there are no significant correlations between the TVA of switchers' current and future teachers for either the within- or across-school TVA estimates.

Since we do not find any evidence that students sort systematically to teachers, we now estimate equation (3) to test whether β_1 is approximately equal to 1. In Figure 2, we plot the event study version of equation (3). We normalize t to be equal to 0 when a student switches schools and regress her mean test scores on

Table 4—Does Future Teacher Value-Added Predict Current Teacher Value-Added When Students Change Schools?

	Coefficient (SE) (1)	Observations (2)
Forward lag of English	0.043 (0.213)	2,046
Forward lag of English (within school)	−0.219 (0.372)	751
Forward lag of math	−0.301 (0.281)	2,046
Forward lag of math (within school)	−0.016 (0.223)	752
Forward lag of Urdu	−0.391 (0.406)	2,046
Forward lag of Urdu (within school)	0.005 (0.262)	752
Forward lag of mean score	−0.189 (0.331)	2,046
Forward lag of mean score (within school)	−0.019 (0.349)	752

Notes: This table tests for bias in the teacher value-added calculations. The current teacher value added of public sector students who change schools in the next period is regressed on the value added of their future teacher. The outcome variables are fixed effects estimates of TVA, while the explanatory variables are empirical Bayes estimates of TVA. When the measure of TVA is across schools instead of within schools, the regressions control for district fixed effects. Observations are at the child-year level, and standard errors are clustered at the teacher level.

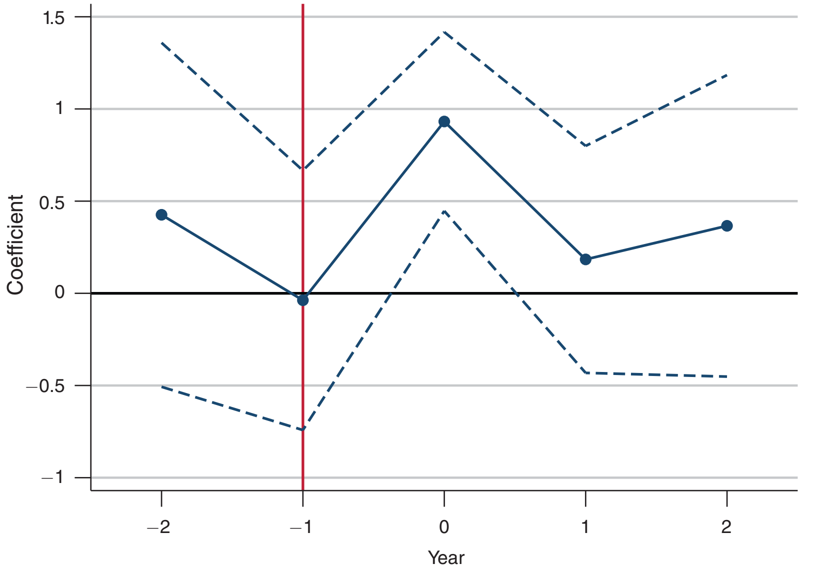

Figure 2. Effect of Mean TVA at $t = 0$ on Switchers' Test Scores over Time

Notes: This figure plots the event study analog of Table 5, which estimates the effect of a student's current teacher's TVA on her test scores after a school switch occurs. The sample consists of public school students who switch schools. $t = 0$ is normalized to be the year a switch occurs, and the figure plots the coefficient of the empirical Bayes measure of mean TVA at $t = 0$ in a regression of the student's mean test score in year t on her teacher's mean TVA in year 0, controlling for fixed effects for the school a student attends in $t = 0$ and the student's lagged mean test scores interacted with grade fixed effects.

Table 5—Out-of-Sample Validation of TVAs in the Public Sector

	Math test score (1)	English test score (2)	Urdu test score (3)	Mean test score (4)
Math TVA	0.710			
	(0.210)			
English TVA		0.836		
		(0.169)		
Urdu TVA			1.185	
			(0.176)	
Mean TVA				1.065
				(0.212)
F-test for coefficient = 1	1.92	0.94	1.11	0.10
School FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Lagged score controls	Y	Y	Y	Y
Observations	1,762	1,762	1,762	1,762
Adjusted R^2	0.615	0.609	0.659	0.703
Clusters	495	495	495	495

Notes: This table tests if TVAs predict the test score gains of school changers who are allocated to the new teachers. If the TVA estimates perfectly predict the "true" teacher value added, these coefficients should be 1. Lagged score controls consist of child class fixed effects and their interactions with lagged test scores. The F-test tests whether the coefficient on the subject TVA in the relevant regression is equal to 1. Standard errors are clustered at the teacher level.

the mean TVA of her teacher in period 0. We again allow for different coefficients by t and control for school fixed effects using the school the student attended at $t = 0$. While the coefficient on TVA at $t = 0$ is approximately 1 (and statistically significant), as we would expect if our TVA estimates are unbiased, the coefficients in the other years are again indistinguishable from 0. Table 5 reports the estimates of equation (3) in the year that the switch occurred and confirms that TVA in a subject is highly predictive of the student's test score gains in that subject. The coefficient is always close to 1, and the formal F-test cannot reject that it is equal to 1 for all three subjects (and the mean across the three), with p-values ranging from 0.17 to 0.76.

Assessing Sampling Variability.—A third exercise assesses the degree to which TVA estimates are sensitive to the sample used for estimation. First, we randomly choose mutually exclusive samples of student test score data with which to estimate the fixed effect version of the TVAs. This allows us to estimate TVAs using data from only rounds two, three, and four separately and from the combinations of rounds two and three, rounds three and four, rounds two and four, and all three rounds. Randomly partitioning the data into mutually exclusive samples ensures that our estimates will not be mechanically correlated because they include the same observations. For each of these datasets, we estimate a different set of mean TVAs.

Table 6 shows that when only one year of data is used, the correlations between the TVA estimates range from 0.03 to 0.40. Adding a second year, however, greatly improves the correlations between the estimates to 0.29 to 0.60. These correlations are similar to the correlations between the two- and three-year estimates, suggesting that adding a third year of data provides less information than the second year. This is not surprising since two years of data allow us to eliminate classroom effects

Table 6—Correlations between TVAs Estimated with Different Public Sector Samples

	Year 2 (1)	Year 3 (2)	Year 4 (3)	Years 2 + 3 (4)	Years 2 + 4 (5)	Years 3 + 4 (6)	Years 2–4 (7)
Year 2	1.000						
Year 3	0.031	1.000					
Year 4	0.396	0.210	1.000				
Years 2 + 3	0.688	0.312	0.267	1.000			
Years 2 + 4	0.672	0.131	0.752	0.409	1.000		
Years 3 + 4	0.321	0.408	0.777	0.290	0.599	1.000	
All years	0.568	0.392	0.674	0.395	0.647	0.627	1.000

Note: This table reports the correlations between fixed effects TVA estimates in the public sector using mutually exclusive, random samples from year 2, year 3, year 4, years 2 and 3, years 2 and 4, years 3 and 4, and years 2–4.

(which are uncorrelated across years) in expectation. In contrast, with only one year of data, the teacher fixed effect will be equivalent to the sum of the teacher effect and the time-varying classroom effect. Online Appendix Figure A9 shows that the distributions of the TVAs become quite similar as we move from using samples from a single year to two or three years.

Multigrade Teaching.—We now assess the effect of multigrade teaching (defined as two or more grades being represented with more than 5 percent of students in a classroom) on the TVA estimates. We assess whether multigrade teaching biases our results in two ways. First, as we discussed in Section IIIC, we directly controlled for multigrade teaching along with other school inputs and student characteristics and re-estimated TVA. As online Appendix Table A6 shows, the TVAs estimated with this richer set of controls are highly correlated with our baseline estimates (correlations of 0.91–0.95 in column 1). Second, we exploit the fact that teachers move in and out of multigrade teaching (251 teachers switch between multi- and single-grade teaching in the public sector) to directly estimate the effect of multigrade classrooms on student test scores, holding the teacher constant. Then, to estimate the effect of multigrade teaching, we include an indicator variable for multigrade in equation (2). We estimate that multigrade teaching is associated with a statistically insignificant 0.037 SD (SE $=$ 0.037) reduction in mean test scores in the public sector, further indicating that multigrade teaching is unlikely to bias our TVA estimates.

Assessing Fade-Out.—In the last subsection, we test the degree to which teacher effects fade out after students have moved on to other teachers. To do so, we re-estimate equation (3), replacing a student's current teacher's TVA with the TVA of the teacher who taught the student one year ago. We exclude the controls for lagged test scores since these are endogenous to lagged TVA. Then, the coefficient on the lagged value of TVA is a measure of how much the effects of being assigned to a specific teacher persist over one year. We also use the same sample of students who have switched schools as before (and exclude these students when we estimate the TVA). By focusing on these switchers, we avoid cases where the estimated coefficient on the lagged TVA is biased by the fact that lagged TVAs are correlated with time-varying school quality or by the fact that a student is taught by the same teacher

Table 7—Estimates of the Persistence of Teacher Effects in the Public Sector

	Mean test score (1)	Mean test score (2)
Mean $TVA_{ij,t-1}$	0.597	0.457
	(0.221)	(0.306)
Measure of TVA	Standard	Leave own class out
Year FE	Y	Y
Grade FE	Y	Y
School FE	Y	Y
F-test for perfect persistence	3.314	3.138
Adjusted R^2	0.383	0.428
Observations	1,512	1,030
Clusters	265	183

Notes: This table reports estimates of the persistence of the effects of TVA on students' test scores from a regression of the students' current test scores on their one-year lagged teachers' mean TVA. The sample is restricted to public school students who have changed schools between the current year and the previous year. Standard errors are clustered at the lagged teacher level. Mean TVAs are cross-school empirical Bayes estimates and calculated leaving out the sample of students who switched schools (in column 1) or all the students in the same classroom as the student i in year t (column 2).

in the current year as in the lagged year. Additionally, we cluster our standard errors at the lagged teacher level. As a robustness check, we test whether our measure of persistence is picking up the persistence of time-varying effects associated with a specific teacher-year combination (e.g., peers) rather than teacher quality. We follow Jacob, Lefgren, and Sims (2010) and re-estimate persistence with a measure of TVA that is estimated leaving out a student's own classroom.[32]

The first column of Table 7 reports our estimate of the persistence of a teacher's effect after one year using our standard empirical Bayes TVA measures. Our estimate suggests that teacher effects are imperfectly persistent in this context, with 60 percent of the initial teacher effect persisting after one year. The remaining column re-estimates the same specification using the alternative measure of TVA that leaves out a student's own classroom. This specification further reduces our sample size and the precision of our results. Nonetheless, our point estimate of 46 percent persistence, though statistically insignificant, is similar to the point estimate that does not exclude own-classroom effects.

Our estimate of one-year persistence is larger than that of Jacob, Lefgren, and Sims (2010), which finds that only 20 percent of a teacher's initial effect persists over one year in the United States. However, our estimate is similar in magnitude to that of Chetty, Friedman, and Rockoff (2014b), which finds that 53 percent of a teacher effect persists over one year. One possible explanation for the difference between our results and the lower estimate by Jacob, Lefgren, and Sims (2010) is that

[32] While it would also be valuable to estimate persistence of TVA's effects over two or three years, in line with Chetty, Friedman, and Rockoff (2014b), our sample sizes limit our ability to do so credibly. Since our cross-school TVA estimates include school effects, estimates of the persistence of two- or three-year lagged TVAs might be upwardly biased by the fact that students spent additional time in the same school (and possibly with the same teacher). However, restricting our sample to students who appear in the current year but switched schools one or two years ago prohibitively decreases our sample size, and we no longer have the statistical power necessary to measure persistence.

Jacob, Lefgren, and Sims (2010) analyzed fade-out on standardized tests designed to assess different skills in different grades. So, their measures of persistence can be interpreted as persistence of a student's *location* in the test score distribution. In contrast, our tests assess (approximately) the same skills every year, and using IRT, we equate tests across years so that all students and years are on the same IRT scale. Therefore, our measures of persistence can be thought of as capturing whether specific skills persist.

IV. Teacher Hiring and Compensation Policies

Having shown that TVA meaningfully measures teacher productivity, we now use our TVA estimates to document three additional results with implications for teacher compensation and hiring policies. First, we re-estimate TVA for *private school teachers* in the sample villages, which allows us to examine wage-TVA relationships in the public and private sectors. We show that wages are unrelated to TVA in the public sector but positively and significantly related to TVA in the private sector. Second, we study a change in the hiring regime in Punjab just prior to our data collection where new teachers were hired on temporary contracts at a significantly lower salary. We show that the TVA of contract teachers in our sample was no lower than that of (higher paid) permanent teachers up to four years after the change in the regime. Taken together, these results suggest that teacher wages in the public sector do not reward more productive teachers and that reducing high public sector premiums does not reduce the quality of the teaching workforce. This leads us, in Section IVC, to simulate the impact of an alternate recruitment policy that hires new contract teachers in each year and selectively tenures the best after a fixed observation period. We discuss the results from this simulation and compare them to alternate policies of teacher reallocation that take advantage of the vast variation in class size in rural Pakistan.

A. *TVA and Wages in the Public and Private Sectors*

Our sample contains information on 1,346 private school teachers linked to 9,741 unique students. We are able to estimate TVAs for 975 of these private school teachers.[33] Online Appendix Tables A7 and A8 show the sources of variation for the TVA calculation in the private sector sample. Online Appendix Table A9 decomposes test score variation into school, classroom, teacher, and student effects for teachers in the private sector following the same strategy as in the public sector. As with public school teachers, private school teachers are key for student learning, although the variance of private sector TVA is lower, with a 1 SD better teacher leading to 0.10 SD higher mean test scores. One plausible explanation for this finding is that private schools are better able to remove poorly performing teachers.

Using the estimates from online Appendix Table A9, we can now construct empirical Bayesian estimates of TVA in the private sector and use these to compare the

[33] As before, the difference in the sample sizes is because children are observed without prior test scores in some years. If a teacher is only observed with students without prior test scores (e.g., if we only observe the teacher in the first round), that teacher is not included in the sample.

Table 8—Relationship between Mean TVA and log Salary for Public and Private School Teachers

	Public				Private	
	log salary (1)	log salary (2)	log salary (3)	log salary (4)	log salary (5)	log salary (6)
Mean TVA		−0.024	−0.083	−0.036	0.918	0.388
		(0.040)	(0.069)	(0.098)	(0.295)	(0.336)
Female	−0.036	−0.035	N.A.	N.A.	−0.413	−0.288
	(0.013)	(0.013)			(0.043)	(0.048)
Local	−0.052	−0.052	−0.049	−0.018	−0.179	−0.040
	(0.019)	(0.019)	(0.032)	(0.044)	(0.029)	(0.035)
Some teacher training	0.518	0.518	0.395	0.845	0.166	0.127
	(0.141)	(0.141)	(0.140)	(0.317)	(0.044)	(0.039)
Has bachelor's or better	0.255	0.255	0.263	0.213	0.326	0.278
	(0.019)	(0.019)	(0.028)	(0.043)	(0.045)	(0.042)
Had >3 years of experience in 2007	0.063	0.064	0.120	0.111	0.024	0.059
	(0.042)	(0.042)	(0.063)	(0.097)	(0.029)	(0.031)
Temporary contract	−0.354	−0.354	−0.327	−0.311		
	(0.032)	(0.032)	(0.059)	(0.091)		
Age	0.058	0.058	0.062	0.039	0.016	0.022
	(0.015)	(0.015)	(0.020)	(0.029)	(0.007)	(0.008)
Age squared	−0.0005	−0.0005	−0.001	−0.0002	−0.0002	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.00008)	(0.00001)
Mean teacher knowledge				0.034		
				(0.028)		
Mean salary	6,987	6,987	6,987	6,745	1,401	1,401
Fixed effects	District	District	School	School	District	School
Adjusted R^2	0.616	0.615	0.662	0.706	0.463	0.769
Observations	1,383	1,383	1,383	919	804	804
F-statistic	108.304	96.587	35.658	14.630	39.188	16.387
Clusters	471	471	471	469	294	294

Notes: This table reports estimates from regressions of log mean teacher salaries in public (columns 1–4) and private (columns 5 and 6) schools on teacher characteristics including mean TVA (columns 2–6), calculated using empirical Bayes and average teacher test scores across subjects (column 4). Mean TVA is measured in student test score standard deviations. The association between female and log salaries in the public sector cannot be credibly estimated in the presence of school fixed effects because the public sector is not coeducational. Very few public schools (29) are observed with both male and female teachers over the course of the sample. All regressions include either district (columns 1, 2, and 5) or school fixed effects (columns 3, 4, and 6), and standard errors are clustered at the school level.

relationship between wages and TVA in the public and private sectors.[34] In Table 8, we regress log salaries on teacher characteristics separately in the public and private sectors. Our estimating equation is

$$\log\left(salary_j\right) \;=\; \beta_0 + \Gamma X_j + \alpha_d + \varepsilon_j,$$

[34] We caution that the sample sizes in the private sector are smaller, which may lead our estimates to be less accurate than in the public sector. Online Appendix Table A8 shows that there are 1,099 classrooms in schools where multiple teachers teach in the private sector (the variation needed to estimate the variance of school effects), while there are 1,914 such classrooms in the public sector. Similarly, online Appendix Table A7 shows that only 332 teachers are observed at least twice (the variation needed to identify the variance of teacher effects) as opposed to 913 teachers in the public sector. In general, our estimates of the school effects in the private sector are much smaller than in the public sector, and in some cases, the covariance between classrooms within a school (particularly for Urdu) are small and negative, which we take to estimate a school effect of zero.

where $\log(salary_j)$ is the log of the mean salary of teacher j, and X_j consists of the same teacher characteristics, as in equation (3). As before, α_d is a district fixed effect, and some specifications (columns 3, 4, and 6) also include school fixed effects. To estimate the degree to which including the mean TVA estimate improves the explanatory power of our regressions, we first exclude mean TVA from the regression (column 1) and then include it in the remaining specifications.

In public schools, controlling for district fixed effects, training is associated with a 52 percent increase in teacher salaries, and having a bachelor's degree is associated with a 26 percent increase.[35] In addition, for a young teacher starting at 22, an additional year of teaching is associated with a 3.7 percent (no school fixed effects) to 4.0 percent (with school fixed effects) increase in wages. Teachers with temporary contracts make 31–35 percent less than teachers with permanent contracts, accounting for differences in other observable characteristics. Two key characteristics associated with higher TVA—the first few years of experience and teacher content knowledge—have no apparent effect on teacher salaries.[36]

In line with findings in India (Muralidharan and Sundararaman 2011), when we add mean TVA to the regressions (columns 2–4), the coefficient is typically small and always negative and insignificant, with no effect on the adjusted R^2. Higher quality teachers are not rewarded with higher salaries in the public sector. This finding would be of limited interest if salaries were entirely determined through a "lock-step" schedule, but in fact, there is substantial room for salaries that reflect performance. The adjusted R^2 after including our extensive controls (column 4) never exceeds 0.71.

We replicate the specifications in columns 2 and 3 for private school teachers in columns 5 and 6. The differences in compensation schemes for the private sector are striking. As has been noted before (Andrabi, Das, and Khwaja 2008), the private sector pays teachers according to their outside option, penalizing women and teachers who are locally resident. The private sector also rewards training and education (in similar ways for education but less so for training). However, the premium on seniority is now much lower, and TVA is highly correlated with salaries. Importantly, a 1 student test score SD increase in TVA is associated with a 92 percent (no school fixed effects) and 39 percent (with school fixed-effects) increase in wages.[37]

We also estimate the relationship between wages and TVA in the private sector using the public sector shrinkage parameters as a lower-bound estimate of this association. This lower-bound estimate indicates that a 1 student test score standard deviation increase in TVA is associated with a highly statistically significant 49 percent increase in wages, and even with this estimate, the difference between the

[35] Almost all public school teachers have at least some training. Therefore, the large association between training and salaries relies on 44 individuals (3 percent of the sample) who have no training.

[36] While we cannot include both experience and age non-parametrically in this Mincerian regression, we can include an indicator variable for whether the teacher had more than three years of experience in the final round for which data was collected. For the most part, we find no additional effect of experience beyond the seniority effect, though when we include school fixed effects, the coefficient on this variable becomes marginally significant.

[37] Using the empirical Bayes measure of mean TVA to correct for attenuation bias is important: our estimates (not shown) indicate that a 1 SD increase in the uncorrected, fixed effect measure of mean TVA would be associated with a (still statistically significant) 13 percent increase in a private sector teacher's salary.

public and private estimates is statistically significant at the 1 percent level. The difference between the estimates with and without school fixed effects suggests that better teachers are sorting into higher-paying, better private schools. But even within schools, higher TVA teachers appear to be paid more, though the estimate is imprecise.

B. *A Case Study of Contract Teachers*

Our data also allow for an interesting case study of what happens to TVA when the government decides to hire contract teachers at much lower salaries. This hiring regime change, described in the working paper version of the paper (Bau and Das 2017), stemmed from worries about low public sector accountability and budgetary concerns due to high teacher wages and benefits. The final precipitating factor was a nuclear test in 1998 that led to international sanctions and greatly worsened Punjab's budgetary position. This led to a hiring freeze, followed by an unprecedented shift from hiring permanent teachers to hiring contract teachers at 35 percent lower wages. Online Appendix Figure A10 shows that while the number of teachers hired each year varies, corresponding to the practice of "batch" hiring in the province (Bari et al. 2013), 1998–2001 was a uniquely long period of low hiring. After normal hiring resumed in 2002, almost all teachers in the province were hired on non-tenured, temporary contracts and received, as we showed in Table 8, 35 percent lower wages than permanent teachers with similar levels of experience.[38] To assess whether the province could hire teachers at lower wages without sacrificing students' learning, we examine whether the TVA of contract teachers hired under this new regime was lower than that of permanent teachers. We caution, however, that we cannot separate pure wage effects from incentive effects due to the contractual status of teachers.

Since not every teacher was hired on a permanent contract before the shock and not every teacher was hired on a temporary contract afterward, we use a fuzzy regression discontinuity design to compare teachers hired right before and after the budgetary shock. This approach is analogous to an instrumental variables regression that incorporates time trends and includes a subset of the sample around the budgetary shock. To exploit the variation induced by the hiring regime change, we estimate the effect of the temporary contract policy (β_1) by instrumenting for an indicator variable for being hired on a temporary contract ($TempContract_j$) with an indicator variable for being hired after 1998 ($Post_j$). The first stage of this two-stage least squares strategy is then

$$\begin{aligned} TempContract_j \; = \; & \delta_0 + \delta_1 Post_j + \delta_2\, month\_hired_j \\ & + \delta_3\, month\_hired_j \times Post_j + \alpha_d + \mu_j, \end{aligned}$$

[38] Contract arrangements in Punjab became more common from 2000–2001 on (Hameed et al. 2014), and in 2004, the government of Punjab announced its Contract Appointment Policy (Cyan 2009).

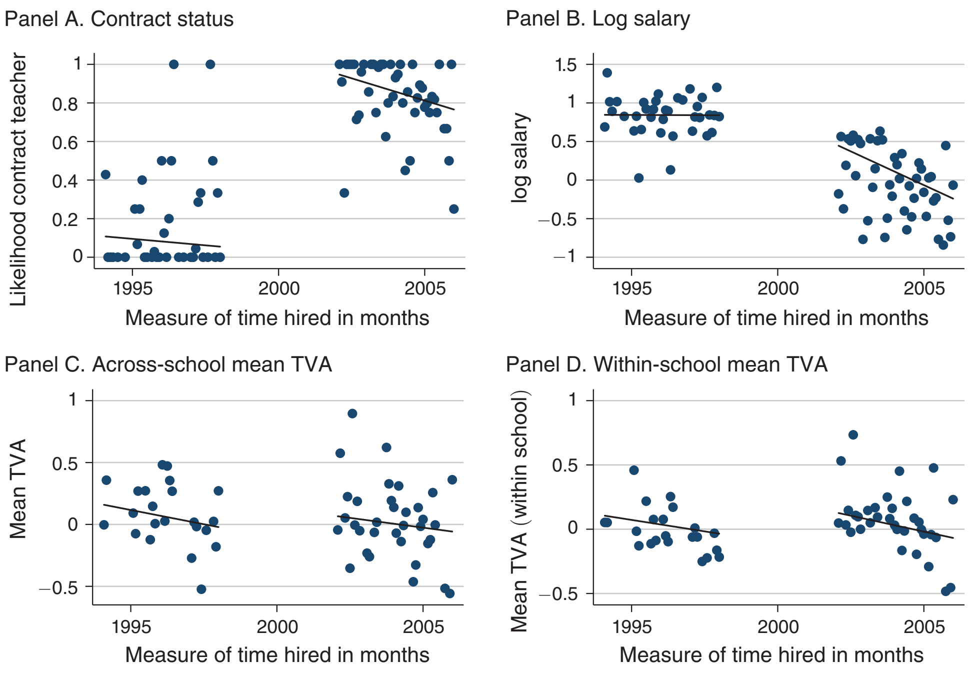

FIGURE 3. TEACHER CONTRACT STATUS, SALARY, AND PRODUCTIVITY BY MONTH HIRED

Notes: This figure plots the trends in contract status, log of mean salary, and cross- and within-school mean TVAs by month hired around the hiring regime change. Each dot is the average value for teachers hired in that month. The lines are the linear best fit and are allowed to have different slopes before and after the regime change. We do not plot the outcomes of the small number of teachers hired during the hiring freeze from 1998–2001, who we also drop from our regression sample.

and the second stage is

$$y_j = \beta_0 + \beta_1 \, TempContract_j + \beta_2 \, month\_hired_j$$
$$+ \beta_3 \, month\_hired_j \times Post_j + \alpha_d + \epsilon_j,$$

where y_j is the outcome of interest, and $month\_hired_j$ is a continuous measure of the month a teacher j was hired. Additionally, we drop the small number of (likely selected) teachers hired during the hiring freeze (from 1998 to 2001) from the sample, and following Lee and Lemieux (2010), we cluster our standard errors at the month-hired level.[39] Furthermore, when our outcome is TVA, we include controls for school facilities, student-teacher ratios, and parents' education and household assets. The key results are the following:

- **Likelihood of contract hiring and effect on salaries:** Figure 3 shows that being hired after 1998 is associated with an 80 percentage point increase in the probability that a teacher is hired on a temporary contract. The second panel

[39] This is the level of discreteness of the forcing variable.

Table 9—Effect of the Discontinuity on Contract Status, Salaries, Teacher Productivity, and Assignment to Schools and Students

	OLS Clusters = 342			RD (Bandwidth = 3 years) Clusters = 65			RD (Bandwidth = 4 years) Clusters = 79		
	Coefficient (1)	SE (2)	Observations (3)	Coefficient (4)	SE (5)	Observations (6)	Coefficient (7)	SE (8)	Observations (9)
First stage	0.810	0.046	3,196	0.833	0.143	868	0.933	0.146	933
log (salary)	−0.284	0.081	3,196	−0.544	0.273	890	−0.444	0.205	959
School basic facilities	0.066	0.057	3,168	0.433	0.270	861	0.274	0.248	925
School extra facilities	−0.480	0.116	3,181	−1.025	0.582	864	−0.922	0.434	929
Student household assets	0.044	0.079	1,487	0.693	0.473	422	0.453	0.372	445
Student mother education	−0.077	0.018	1,487	0.020	0.129	422	−0.001	0.116	445
Student father education	−0.042	0.023	1,487	−0.130	0.094	422	−0.134	0.081	445
Across-school mean TVA	0.011	0.039	1,273	0.293	0.262	345	0.261	0.201	359
Within-school mean TVA	0.042	0.027	1,203	0.353	0.179	320	0.288	0.134	334

Notes: This table presents OLS and fuzzy regression discontinuity results for the effect of temporary contracts on teacher characteristics. The first row reports the first-stage estimates of the effect of being hired after 1998 on temporary contract status. The remaining cells report the coefficients (and standard errors and sample sizes) for regressions of the row variables on temporary contract status in the OLS (columns 1–3) or second-stage IV (columns 4–9) regressions. All regressions include controls for month hired and its interaction with being hired after 1998. The RD samples include either teachers hired three years before 1998 and three years after 2001 or teachers hired four years before 1998 and four years after 2001. The *F*-tests for the effect of being hired after 1998 on temporary contract status are 308.98 (full sample), 33.78 (three-year bandwidth), and 40.69 (four-year bandwidth), respectively. Characteristics that are observed at the teacher-year or student-year level were normalized to create time-invariant teacher-level indices by calculating the teacher fixed effect, controlling for year fixed effects, and then de-meaned at the district level. The household asset measure is the first factor of a principal components analysis of indicator variables for ownership of beds, a radio, a television, a refrigerator, a bicycle, a plow, agricultural tools, tables, fans, a tractor, cattle, goats, chickens, watches, a motor rickshaw, a scooter, a car, a telephone, and a tube well following methods discussed by Filmer and Pritchett (2001). The two indices for school facilities are constructed as the first predicted component from principal components analyses of indicator variables for "basic" and "extra" school facilities. Extra school facilities consist of a library, computer, sports, a hall, school wall, fans, and electricity. The basic facilities consist of whether the school has desks/chairs as its seating arrangement, blackboards per child, toilets per child, and classrooms per child. TVAs are measured in student-level test score standard deviations. Standard errors are clustered at the month-hired level.

shows the discontinuity in salaries, with regression equivalents in Table 9. The salary declines range from 28 percent (OLS) to 44–54 percent (fuzzy RD), compared to 35 percent in the Mincerian regression, which accounts for only the observable characteristics of teachers.

- **Allocation of contract teachers to schools and students:** Contract teachers joined schools with significantly fewer extra facilities (Table 9). However, as online Appendix Table A10 shows, there was no differential pre-trend in test scores for (i) schools that received a contract teacher compared to those that did not (column 1) or (ii) students within schools who received a contract teacher in the future and those who did not (column 2). Finally, a student's historical average test score gains prior to being assigned to a contract teacher do not predict assignment to a contract teacher (column 3). Overall, we find little evidence that students selectively sort to contract teachers.

- **TVA of contract versus permanent teachers:** The results from the RD design (Table 9) suggest that the TVA of contract teachers was higher by approximately 0.3 (student test score) SD compared to permanent teachers, although this is imprecisely estimated. Online Appendix Table A11 reports the results of the fuzzy regression discontinuities with bandwidths up to 10 years. Regardless of whether we use larger bandwidths or compare teachers within the same school, we never estimate a substantial negative contract teacher effect.[40] Therefore, while there isn't conclusive evidence that the contract policy raised teacher quality, it is very unlikely that the policy *lowered* teacher productivity.[41]

- **Later versus immediate entrants:** The RD strategy suggests that applicants hired immediately after the budgetary shock have a similar TVA to permanent teachers, but the quality of applicants may change over time.[42] To test whether this is the case, we compare the test scores of students taught by contract teachers hired in later rounds to those hired in earlier rounds. To mitigate experience effects, which lead observed teacher quality to decrease for teachers who are hired later, we only compare the outcomes of the contract teachers' students when we see them with inexperienced teachers (teachers with zero or one year of experience). We also include the permanent teachers hired before 1998 in our regression sample as a control group that allows us to account for testing round fixed effects. For this sample of inexperienced contract teachers and permanent teachers, we regress student test scores on lagged student test scores (allowing the effect to vary by grade), district and year fixed effects, a continuous measure of the month the teacher is hired, an indicator variable for a teacher being hired after 1998, and the interaction between the continuous measure of month hired and the indicator variable for being hired after 1998. If teacher quality is declining after the policy, we expect this interaction to be negative and significant. We find that the interaction term is small and insignificant (a coefficient of -0.007 with a standard error of 0.024). Thus, we find no evidence that teacher quality decreased in response to decreased teacher salaries and teacher tenure for later hires.

In Bau and Das (2017), we provide additional robustness tests, showing that (i) the estimated effect of temporary contract status on mean TVA is very similar when the TVA estimation includes controls for average socioeconomic status and lagged test scores (following Altonji and Mansfield 2018), and (ii) that there is no differential attrition from the sample for contract teachers or their students. We also provide suggestive evidence that the observable characteristics of the new teachers

[40] Larger bandwidths include contract teachers with lower experience levels, which is likely to negatively bias our estimates of the effect of temporary contract teachers on test scores.

[41] To ensure that our TVA results are robust, we also adjust the analytical standard errors for estimation error in TVA by estimating the RD p-values with a clustered bootstrap procedure (see online Appendix Table A11). The pattern of significance for the estimates is similar to the pattern obtained using analytic standard errors, although the within-school effect for the three-year bandwidth is now only marginally statistically significant.

[42] If teachers invest in the profession by incurring a large fixed cost, the longer term effects of the policy may be different from the immediate effects. Primary school teachers in Pakistan undertake a one-year primary school teacher-training program, while secondary school teachers are required to complete either a one-year program leading to the bachelor of education diploma or a three-year program leading to a bachelor of education degree. We can observe TVA up to four years after the change in regime, extending our hiring results to teachers who had not yet made the decision to invest in teacher-specific skills prior to the regime change.

did not differ from those of the permanent teachers. Finally, we show that, with a conservative discount rate of 3 percent, even if contract teachers are made permanent after 10 years (as they were in 2012 in our case), these temporary differences in wages correspond to a 13 percent decline in permanent wages. Even at this low discount rate, the contract teacher policy substantially reduced permanent wages for the teachers hired without negatively affecting incoming teacher quality.

C. *Teacher Hiring Policy Simulation*

Our estimates of the variance of mean TVA in Table 2 suggest that teacher quality matters for students' outcomes. Additionally, our findings in the previous two subsections suggest that current teacher hiring and compensation policies are unlikely to produce the best outcomes for students. Thus, a policy where contract teachers are observed for a probationary period before being confirmed/tenured, such as the one proposed by Staiger and Rockoff (2010) in the United States, as well as Pritchett and Murgai (2006) and Muralidharan (2016), may increase student test scores. Indeed, if the bottom 5 percent of teachers could be perfectly identified and replaced with teachers with average TVA, test scores of affected students would increase by 0.31 SD. Of course, because the fraction of students affected by this policy is necessarily small, the impact on average student test scores is a lower 0.02 SD ($= 0.31/20$).

To better understand the effects of these types of policies, we now simulate the impact of selective tenure policies. In our simulations, some teachers retire each year and must be replaced. New contract teachers are hired and observed for either one, two, or five years before being confirmed, and we simulate what happens when the bottom 5, 10, 50, 60, 70, and 80 percent of teachers according to their observed TVA are not confirmed. We use our estimates of the variance of teacher, classroom, and idiosyncratic student effects to simulate test scores and calculate teacher's observed value-added at the end of the probationary period.[43] We then replace teachers below the cutoff with new teachers drawn from the mean TVA distribution and subtract 0.08 SD from the TVA of teachers with only zero or one years of experience (consistent with our experience effects in Table 3). Online Appendix D provides further details.

In our simulations, greater selectivity affects TVA through three forces. As selectivity increases, so does the mean TVA of the tenured pool. However, the fraction of pre-tenure teachers also has to increase to keep the size of the teaching workforce constant. Since newly hired contract teachers are randomly drawn from the pool of applicants, their TVA will be the mean TVA of the applicant pool discounted for the experience effect. Finally, the longer the period of observation, the smaller the errors in the tenure process but the greater the number of years that lower performing contract teachers remain in the workforce.

In Table 10, we report the effects of each policy on mean TVA (and therefore, student test scores) after 2, 5, and 15 years, and in the steady state. Since the specific

[43] We ignore school effects since these would be perfectly identified to the policymaker with enough years of data.

Table 10—Mean Student Test Score Gains in Simulated Hiring Policies

Percent not tenured (1)	2 years (2)	5 years (3)	15 years (4)	Steady state (5)
Panel A. Confirmation after 1 year				
5	(−0.01, 0)	(−0.01, 0)	(0, 0)	0.01
10	(−0.01, 0)	(0.001, 0)	(0, 0.01)	0.01
50	(−0.01, −0.01)	(0, 0)	(0.02, 0.02)	0.04
60	(−0.01, −0.01)	(0, 0)	(0.02, 0.02)	0.05
70	(−0.01, −0.01)	(0, 0)	(0.02, 0.02)	0.06
80	(−0.01, −0.01)	(0, 0)	(0.02, 0.03)	0.06
Panel B. Confirmation after 2 years				
5	(−0.01, 0)	(−0.01, 0)	(0, 0.01)	0.01
10	(−0.01, −0.01)	(−0.01, 0)	(0, 0.01)	0.01
50	(−0.01, −0.01)	(0, 0)	(0.02, 0.02)	0.05
60	(−0.01, −0.01)	(0, 0)	(0.02, 0.02)	0.06
70	(−0.01, −0.01)	(−0.01, 0)	(0.02, 0.03)	0.06
80	(−0.02, −0.02)	(−0.01, −0.01)	(0.02, 0.02)	0.07
Panel C. Confirmation after 5 years				
5	(−0.01, 0)	(−0.01, 0)	(−0.01, 0.01)	0.01
10	(−0.01, 0)	(−0.01, 0)	(0, 0.01)	0.02
50	(−0.01, 0)	(0, 0)	(0.02, 0.02)	0.06
60	(0, 0)	(0, 0)	(0.02, 0.03)	0.07
70	(0, 0)	(0, 0)	(0.02, 0.03)	0.07
80	(0, 0)	(0, 0)	(0.02, 0.02)	0.06

Notes: This table reports mean steady-state student test score gains from policies that replace the observed worst 5, 10, 50, 60, 70, or 80 percent of teachers after 1, 2, or 5 years. The mean test score gains were calculated by simulating student test scores and teacher hiring and firing under each policy until steady state was achieved using the estimates of teacher and class effects and student idiosyncratic shocks in Table 2. To obtain 95 percent confidence intervals for the effects of the policy prior to achieving the steady state (reported in columns 2, 5, and 15), we simulated the policy 200 times and report the values at the 2.5 and 97.5 percentiles of the simulated policy. Online Appendix D describes the policy simulations in detail.

draws of teachers in year t might affect the reported effects in a given year, we simulate the policy 200 times and report the 95 percent confidence intervals for the effects at $t = 2, 5, 15$. In our estimates, there is little gain from waiting more than two years to learn more about a teacher's TVA. In general, the selectivity of the tenure process has a larger effect compared to the length of the observation period. Achieving higher average gains (in the range of a 0.05–0.07 SD increase in the average teacher's TVA) requires not confirming a large percent of teachers. This is in line with Staiger and Rockoff (2010), which shows that not confirming the bottom 80 percent of teachers would maximize test score gains in the United States. When 80 percent of teachers are not confirmed, the average yearly effect of the policy on test scores is 0.07 SD, compared to 0.08 SD for the United States. Given the estimated persistence of teacher effects, this would raise average test scores by 0.17 SD at the end of 10 years of schooling. But, if the implementation of the policies resembled the simulations, only one half of the steady-state gains will be achieved in the first 15 years of the policy.

Thus, the policy of employing contract teachers can lead to substantial fiscal savings, but whether the learning gains alone justify a large policy change remains an open question. We note that in our sample, there are alternative meaningful policy changes that exploit the vast variation in classroom size. For instance, we can also simulate the results of a policy that reallocates the highest performing teachers to the largest classrooms so that more students can benefit from higher quality teachers. This is a policy-relevant simulation since transfers are centrally managed and fairly frequent in our context. We simulate this policy under the assumption that class size effects are separable from teacher quality, and we keep the class sizes the same as before the teachers are reallocated. We find that reallocating teachers in this way within the same village generates average mean test score gains of 0.025 SD, while reallocating teachers across villages generates test score gains of 0.053 SD. These are close to the gains from a selective tenure policy and may be a more politically palatable alternative.

V. Conclusion

This paper provides among the first estimates of TVA, its dispersion, and its correlation with observable teacher characteristics from a low-income country. We are able to establish the validity of TVA models in this context, despite additional complications from only observing small schools and classrooms. Like in the United States, effective teachers raise student test scores, but what makes a teacher effective is only weakly linked to observed teacher characteristics. Unlike in the United States, teacher content knowledge is associated with higher TVA. This raises important questions about how teachers should be recruited and the relative benefits of systems with a probationary period followed by tenure for high performers. We simulate the impacts of such a policy and show that at high levels of selectivity, this policy could increase yearly test scores by 0.07 SD in steady state.

We also demonstrate the weak link between TVA and wages in the public sector, even though there is a very high correlation between TVA and wages in the private sector. Like Duflo, Dupas, and Kremer (2015) and Muralidharan and Sundararaman (2013), we find that contract teachers with lower wages have equal or greater TVA relative to permanent teachers. This suggests that there is severe misallocation of wages and TVA in the public sector and raises further questions about public sector wage determination in low-income countries. If teacher compensation systems should be designed to recruit and retain high-quality teachers at the lowest costs and incentivize teacher effort, our results suggest that there is significant room for the better design of public sector compensation regimes.

REFERENCES

Altonji, Joseph G., and Richard K. Mansfield. 2018. "Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects." *American Economic Review* 108 (10): 2902–46.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2008. "A Dime a Day: The Possibilities and Limits of Private Schooling in Pakistan." *Comparative Education Review* 52 (3): 329–55.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, and Tristan Zajonc. 2008. *Learning and Educational Achievements in Punjab Schools* (*LEAPS*)*: Insights to Inform the Education Policy Debate*. Washington, DC: World Bank.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2006. "Religious School Enrollment in Pakistan: A Look at the Data." *Comparative Education Review* 50 (3): 446–77.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3 (3): 29–54.

Antecol, Heather, Ozkan Eren, and Serkan Ozbeklik. 2015. "The Effect of Teacher Gender on Student Achievement in Primary School." *Journal of Labor Economics* 33 (1): 63–89.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131 (3): 1415–53.

Aucejo, Esteban M. 2011. "Assessing the Role of Teacher-Student Interactions." http://econ.sciences-po.fr/sites/default/files/esteban_aucejo.pdf.

Bari, Faisal, Reehana Raza, Monazza Aslam, Bisma Khan, and Neelum Maqsood. 2013. *An Investigation into Teacher Recruitment and Retention in Punjab*. Lahore, Pakistan: Institute of Development and Economic Alternatives (IDEAS).

Bau, Natalie. 2019. "Estimating an Equilibrium Model of Horizontal Competition in Education." CEPR Working Paper DP13924.

Bau, Natalie, and Jishnu Das. 2017. "The Misallocation of Pay and Productivity in the Public Sector: Evidence from the Labor Market for Teachers." World Bank Working Paper 8050.

Bietenbeck, Jan, Marc Piopiunik, and Simon Wiederhold. 2018. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" *Journal of Human Resources* 53: 553–78.

Blum, Nicole, and Rashmi Diwan. 2007. *Small, Multigrade Schools and Increasing Access to Primary Education in India: National Context and NGO Initiatives. CREATE Pathways to Access. Research Monograph No. 17*. Washington, DC: Education Resources Information Center (ERIC).

Bold, Tessa, Deon P. Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian W. Stacy, Jakob Svensson, and Waly Wane. 2016. "What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa." World Bank Working Paper 7956.

Bruns, Barbara, Alain Mingat, and Ramahatra Rakotomalala. 2003. *Achieving Universal Primary Education by 2015: A Chance for Every Child*. Washington, DC: World Bank Publications.

Buhl-Wiggers, Julie, Jason T. Kerwin, Jeffrey A. Smith, and Rebecca Thornton. 2017. "The Impact of Teacher Effectiveness on Student Learning in Africa." https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=CSAE2017&paper_id=1008.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20 (1): 91–116.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2015. "Response to Rothstein (2014) 'Revisiting the Impacts of Teachers.'" CEPR Discussion Paper 10768.

Cyan, Musharraf. 2009. *Contract Employment Policy Review*. Punjab: Punjab Government Efficiency Improvement Program.

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan. 2016. "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India." *American Economic Review* 106 (12): 3765–99.

Das, Jishnu, and Tristan Zajonc. 2010. "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement." *Journal of Development Economics* 92 (2): 175–87.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. "Double for Nothing? The Effects of Unconditional Teacher Salary Increases on Student Performance." *Quarterly Journal of Economics* 133 (2): 993–1039.

Dee, Thomas S. 2007. "Teachers and the Gender Gaps in Student Achievement." *Journal of Human Resources* 42 (3): 528–54.

Disney, Richard, and Amanda Gosling. 1998. "Does It Pay to Work in the Public Sector?" *Fiscal Studies* 19 (4): 347–74.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–74.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." *Journal of Public Economics* 123: 92–110.

Dustmann, Christian, and Arthur van Soest. 1998. "Public and Private Sector Wages of Male Workers in Germany." *European Economic Review* 42 (8): 1417–41.

Filmer, Deon, and Lant H. Pritchett. 2001. "Estimating Wealth Effects without Expenditure Data or Tears: An Application to Educational Enrollments in States of India." *Demography* 38 (1): 115–32.

Hameed, Yousef M.Y., Rama M. Dilshad, Muhammad A. Malik, and Humera Batool. 2014. "Comparison of Academic Performance of Regular and Contract Teachers at Elementary Schools." *Asian Journal of Management Sciences and Education* 3 (1): 89–95.

Harris, Douglas N., and Tim R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." https://pdfs.semanticscholar.org/0304/3bcd2adaf017f85a7fafb082b4bab0d77eeb.pdf.

Hoffmann, Florian, and Philip Oreopoulos. 2009. "A Professor Like Me: The Influence of Instructor Gender on College Achievement." *Journal of Human Resources* 44 (2): 479–94.

Idara-e-Taleem-o-Aagahi. 2013. *Status of Teachers in Pakistan, 2013*. Lahore: UNESCO and ITA.

Ishtiaq, Nohman. 2012. *Understanding Punjab Education Budget 2012–2013: A Brief for Standing Committee on Education, Provincial Assembly of the Punjab*. Islamabad and Lahore: Pakistan Institute of Legislative Development and Transparency.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45 (4): 915–43.

Jimenez, Emmanuel, Marlaine E. Lockheed, and Vicente Paqueo. 1991. "The Relative Efficiency of Private and Public Schools in Developing Countries." *World Bank Research Observer* 6 (2): 205–18.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

Lucifora, Claudio, and Dominique Meurs. 2006. "The Public Sector Pay Gap in France, Great Britain and Italy." *Review of Income and Wealth* 52 (1): 43–59.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education* 4 (4): 572–606.

Metzler, Johannes, and Ludger Woessmann. 2012. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence From Within-Teacher Within-Student Variation." *Journal of Development Economics* 99 (2): 486–96.

Miller, Raegen. 2012. *Teacher Absence as a Leading Indicator of Student Achievement: New National Data Offer Opportunity to Examine Cost of Teacher Absence Relative to Learning Loss*. Washington, DC: Education Resources Information Center (ERIC).

Mulkeen, Aidan G., and Cathal Higgins. 2009. *Multigrade Teaching in Sub-Saharan Africa: Lessons from Uganda, Senegal, and the Gambia.* Washington, DC: World Bank.

Muralidharan, Karthik. 2016. "A New Approach to Public Sector Hiring in India for Improved Service Delivery." *India Policy Forum* 12: 187–236.

Muralidharan, Karthik, and Michael Kremer. 2008. "Chapter 5—Public and Private Schools in Rural India." In *School Choice International: Exploring Public-Private Partnerships*, edited by Rajashri Chakrakbarti and Paul E. Peterson, 91–110. Cambridge, MA: MIT Press.

Muralidharan, Karthik, and Ketki Sheth. 2016. "Bridging Education Gender Gaps in Developing Countries: The Role of Female Teachers." *Journal of Human Resources* 51 (2): 269–97.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." NBER Working Paper 19440.

Pritchett, Lant, and Rinku Murgai. 2006. "Teacher Compensation: Can Decentralization to Local Bodies Take India from the Perfect Storm through Troubled Waters to Clear Sailing?" *India Policy Forum* 3 (1): 123–77.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–52.

Singh, Abhijeet. 2015. "Private School Effects in Urban and Rural India: Panel Estimates at Primary and Secondary School Ages." *Journal of Development Economics* 113: 16–32.

Singh, Abhijeet. 2019. "Learning More with Every Year: School Year Productivity and International Learning Divergence." *Journal of the European Economic Association,* JVZ033.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24 (3): 97–118.

UNESCO Islamabad. 2013. *Education Budgets: A Study of Selected Districts of Pakistan*. Pakistan: UNESCO Islamabad.

Wiswall, Matthew. 2013. "The Dynamics of Teacher Quality." *Journal of Public Economics* 100: 61–78.