The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation*

Johannes Metzler† and Ludger Woessmann‡

Abstract

Teachers differ greatly in how much they teach their students, but little is known about which teacher attributes account for this. We estimate the causal effect of teacher subject knowledge on student achievement using within-teacher within-student variation, exploiting a unique Peruvian 6th-grade dataset that tested both students and their teachers in two subjects. We circumvent omitted-variable and selection biases using student and teacher fixed effects and observing teachers teaching both subjects in one-classroom-per-grade schools. After measurement-error correction, one standard deviation in subject-specific teacher achievement increases student achievement by about 10 percent of a standard deviation.

Keywords: teacher knowledge, student achievement, Peru
JEL classification: I20, O15

October 13, 2010

* We thank Sandy Black, Behrman, David Card, Paul Glewwe, Eric Hanushek, Patrick Kline, Michael Kremer, Javier Luque, Lant Pritchett, Martin West, and seminar participants at UC Berkeley, Harvard University, the University of Munich, the Ifo Institute, the Glasgow congress of the European Economic Association, and the CESifo Area Conference on Economics of Education for helpful discussion and comments. Special gratitude goes to Andrés Burga León of the Ministerio de Educación del Perú for providing us with reliability statistics for the teacher tests. Woessmann is thankful for the support and hospitality by the W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship of the Hoover Institution, Stanford University. Support by the Pact for Research and Innovation of the Leibniz Association and the Regional Bureau for Latin America and the Caribbean of the United Nations Development Programme are also gratefully acknowledged.

† University of Munich and Monitor Group. jo.metzler@gmail.com.

‡ University of Munich and Ifo Institute for Economic Research, Poschingerstr. 5, 81679 Munich, Germany. CESifo and IZA. woessmann@ifo.de. www.cesifo.de/woessmann.

I. Introduction

One of the biggest puzzles in educational production today is the teacher quality puzzle: While there is clear evidence that teacher quality is a key determinant of student learning, little is known about which specific observable characteristics of teachers can account for this impact (e.g., Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007). In particular, there is little evidence that those characteristics most often used in hiring and salary decisions, namely teachers' education and experience, are crucial for teacher quality. Virtually the only attribute that has been shown to be more frequently significantly correlated with student achievement is teachers' academic skills measured by scores on achievement tests (cf. Wayne and Youngs 2003; Eide, Goldhaber, and Brewer 2004; Hanushek and Rivkin 2006). The problem with the latter evidence, however, is that issues of omitted variables and non-random selection are very hard to address when estimating causal effects of teacher characteristics. In this paper, we provide estimates of the impact of teachers' academic skills on students' academic achievement that circumvent problems from unobserved teacher traits and non-random sorting of students to teachers.

We use a unique primary-school dataset from Peru that contains test scores in two academic subjects not only for each student, but also for each teacher. This allows us to identify the impact of teachers' academic performance in a specific subject on students' academic performance in the subject, while at the same time holding constant any student characteristics and any teacher characteristics that do not differ across subjects. We can observe whether the same student taught by the same teacher in two different academic subjects performs better in one of the subjects if the teacher's knowledge is relatively better in this subject. Thus, our model can identify the effect based on within-teacher within-student variation by controlling for student fixed effects, teacher fixed effects, and subject fixed effects. We can additionally restrict our analysis to small, mostly rural schools with at most one teacher per grade, excluding any remaining possibility that parents may have chosen a specific teacher for their students in both subjects based on the teacher's specific knowledge in one subject, or of bias from non-random classroom assignments more generally (cf. Rothstein 2010).

We find that teacher subject knowledge exerts a statistically and quantitatively significant impact on student achievement. Once corrected for measurement error in the teacher test-score measure, our results suggest that a one standard deviation increase in teacher test scores raises

student test scores by about 10 percent of a standard deviation. This means that if a student was moved, for example, from a teacher at the 5th percentile of the distribution of teacher subject knowledge to a teacher at the median of this distribution, the student's achievement would increase by 16.6 percent of a standard deviation by the end of the school year. Thus, teacher subject knowledge is a relevant observable factor that is part of overall teacher quality.

Methodologically, our identification strategy for teacher effects extends the within-student comparisons in two different subjects proposed by Dee (2005; 2007) as a way of holding constant student characteristics.[1] His approach, however, uses variation across different teachers, as the analyzed teacher characteristics – gender, race, and ethnicity – do not vary within teachers. As a consequence, the identifying variation may still be related to issues of selection and unobserved teacher characteristics. By contrast, our approach uses variation not only within individual students but also within individual teachers. The effect of interest is still identified because subject knowledge does vary within individual teachers and because our dataset allows observing teachers' knowledge in two different subjects. Based on correlated random effects models, we also confirm that the overidentification restrictions implied in the fixed-effects model – that teacher knowledge effects and selection effects are the same across subjects – are valid.

The vast existing literature on education production functions hints at teacher knowledge as one – possibly the only – factor reasonably consistently associated with growth in student achievement.[2] In his early review of the literature, Hanushek (1986, p. 1164) concluded that "the closest thing to a consistent finding among the studies is that 'smarter' teachers, ones who perform well on verbal ability tests, do better in the classroom" – although even on this account, the existing evidence is not overwhelming.[3] Important studies estimating the association of teacher test scores with student achievement gains in the United States include Hanushek (1971; 1992), Summers and Wolfe (1977), Murnane and Phillips (1981), Ehrenberg and Brewer (1995),

[1] Further examples using this between-subject method to identify effects of specific teacher attributes such as gender, ethnicity, and credentials include Ammermüller and Dolton (2006) and Clotfelter, Ladd, and Vigdor (2007).

[2] The Coleman et al. (1966) report, which initiated this literature, first found that verbal skills of teachers were associated with better student achievement.

[3] A decade later, Hanushek (1997, p. 144) counts a total of 41 estimates of the effect of teacher test scores and finds that "of all the explicit measures [of teachers and schools] that lend themselves to tabulation, stronger teacher test scores are most consistently related to higher student achievement." Similarly, Eide, Goldhaber, and Brewer (2004, p. 233) suggest that, compared to more standard measures of teacher attributes, "a stronger case can be made for measures of teachers' academic performance or skill as predictors of teachers' effectiveness." Hanushek and Rivkin (2006) provide a similar recent assessment of the literature.

Ferguson and Ladd (1996), Rowan, Chiang, and Miller (1997), Ferguson (1998), and Rockoff et al. (2008). Examples of education production function estimates with teacher test scores in developing countries include Harbison and Hanushek (1992) in rural Northeast Brazil, Tan, Lane, and Coustère (1997) in the Philippines, Bedi and Marshall (2002) in Honduras, and Behrman, Ross, and Sabot (2008) in rural Pakistan.[4]

However, the existing evidence on the association between teacher knowledge and student achievement – be it in level or value-added form – is still likely to suffer from bias due to unobserved student characteristics, omitted school and teacher variables, and non-random sorting and selection into classrooms and schools (cf. Glewwe and Kremer 2006). Obvious examples where such bias would occur include incidents where better-motivated teachers incite more student learning but also accrue more subject knowledge; where parents with a high preference for educational achievement both choose schools or classrooms within schools with teachers of higher subject knowledge and also further their children's learning in other ways; and where principals place students with higher learning gains into classrooms of teachers with higher knowledge. In short, estimates of the effect of teacher knowledge (or most other observable teacher characteristics) that convincingly overcome such biases from omitted variables and selection are so far missing in the literature.

Apart from omitted variables and selection bias, an additional serious problem when estimating teacher effects is measurement error, which is likely to attenuate estimated effects (cf. Glewwe and Kremer 2006, p. 989). A first source of measurement error arises if the available measure proxies only poorly for the concept of teachers' subject knowledge. In most existing studies, the tested teacher skills can be only weakly tied to the academic knowledge in the subject in which student achievement is examined and cannot distinguish between subject-specific cognitive and general non-cognitive teacher skills (see Rockoff et al. 2008 for a notable exception). Often, the examined skill is not subject-specific, either when based on a rudimentary measure of verbal ability (e.g., Coleman et al. 1966) or when aggregated across a range of skills including verbal and pedagogic ability as well as knowledge in different subjects (e.g., Summers and Wolfe 1977). A second source of measurement error arises because any specific test will measure teacher subject knowledge only with considerable noise. This is most evident when the

[4] See Glewwe and Kremer (2006) and Behrman (2010) for recent reviews of the literature on education production functions in developing countries.

result of one single math question is used as an indicator of math skills (cf. Rowan, Chiang, and Miller 1997), but it applies more generally because the reliability of any given item battery is not perfect, thus giving rise to attenuation bias when estimating effects of teacher knowledge. In this paper, we have access to separate tests of teachers' subject-specific knowledge in math and reading that are each generated from batteries of test questions using psychometric modeling. These provide not only separate measures of teacher subject knowledge, but also consistent estimates of the internal reliability metric of the tests based on splitting the underlying test items in halves, which allow for adjustment for biases due to measurement error based on classical measurement error theory.

Understanding the causes of student achievement is important not least because of its substantial economic importance. For example, differences in educational performance – when measured by student achievement tests – can account for an important part of the difference in long-run economic growth between Latin America and other world regions, as well as within Latin America (Hanushek and Woessmann 2009). Peru has been doing comparatively well in terms of school attendance, but not in terms of educational quality as measured by achievement tests (cf. World Bank 2007; Luque 2008). Its average performance on international student achievement tests is dismal compared to developed countries (cf. OECD 2003) and just below the average of other Latin American countries. Of the 16 countries participating in a Latin American comparative study in 2006, Peruvian 6th-graders ranked 9 and 10, respectively, in math and reading (LLECE 2008). Policy-makers in developing countries may be well advised to place specific focus on teachers' subject knowledge in teacher recruitment, training, and compensation policies.

In what follows, Section II derives our econometric identification strategy. Section III describes the dataset that allows its implementation and reports descriptive statistics. Section IV presents our results and robustness tests, and the final section concludes.

II. Empirical Identification

We consider an education production function with an explicit focus on teacher skills:

(1a) $$y_{i1} = \beta_1 T_{t1} + \gamma U_{t1} + \alpha Z_i + \delta X_{i1} + \mu_i + \tau_{t1} + \varepsilon_{i1}$$

(1b) $$y_{i2} = \beta_2 T_{t2} + \gamma U_{t2} + \alpha Z_i + \delta X_{i2} + \mu_i + \tau_{t2} + \varepsilon_{i2}$$

where y_{i1} and y_{i2} are test scores of student i in subjects 1 and 2 (math and reading in our application below), respectively. Teachers t are characterized by subject-specific teacher knowledge T_{tj} and non-subject-specific teacher characteristics U_t such as pedagogical skills and motivation. (The latter differ across the two equations only if the student is taught by different teachers in the two subjects.) Additional factors are non-subject-specific (Z_i) and subject-specific (X_{ij}) characteristics of students and schools. The error term is composed of a student-specific component μ_i, a teacher-specific component τ_t, and a subject-specific component ε_{ij}.

The coefficient vectors β_1, β_2, and γ characterize the impact of all subject-specific and non-subject-specific teacher characteristics that constitute the overall teacher quality effect as estimated by value-added studies (cf. Hanushek and Rivkin 2010). As indicated earlier, several sources of endogeneity bias, stemming from unobserved student and teacher characteristics or from non-random sorting of students to schools and teachers, are likely to hamper identification of the effect of teacher subject knowledge in equations (1a) and (1b).

Following Chamberlain (1982), we model the potential correlation of the unobserved student effect μ_i with the observed inputs in a general setup:

$$\mu_i = \eta_1 T_{t1} + \eta_2 T_{t2} + \theta_1 U_{t1} + \theta_2 U_{t2} + \chi Z_i + \phi X_{i1} + \phi X_{i2} + \omega_i \tag{2}$$

where ω_i is uncorrelated with the observed inputs. We allow the η and θ parameters to differ across subjects, but assume that the parameters on the subject-specific student characteristics are the same across subjects.

Substituting equation (2) into equations (1a) and (1b) and collecting terms yields the following correlated random effects models:[5]

$$y_{i1} = (\beta_1 + \eta_1) T_{t1} + \eta_2 T_{t2} + (\gamma + \theta_1) U_{t1} + \theta_2 U_{t2} + (\alpha + \chi) Z_i + (\delta + \phi) X_{i1} + \phi X_{i2} + \tau_{t1} + \varepsilon'_{i1} \tag{3a}$$

$$y_{i2} = \eta_1 T_{t1} + (\beta_2 + \eta_2) T_{t2} + \theta_1 U_{t1} + (\gamma + \theta_2) U_{t2} + (\alpha + \chi) Z_i + \phi X_{i1} + (\delta + \phi) X_{i2} + \tau_{t2} + \varepsilon'_{i2} \tag{3b}$$

where $\varepsilon'_{ij} = \varepsilon_{ij} + \omega_i$. In this model, teacher scores in each subject enter the reduced-form equation of both subjects. The η coefficients capture the extent to which standard models would

[5] This correlated random effects specification is similar to siblings models of earnings effects of education (Ashenfelter and Krueger 1994; Card 1999). We thank David Card for alerting us to this interpretation of our model.

be biased due to the omission of unobserved teacher factors, while the β parameters represent the structural effect of teacher subject knowledge.

The model setup with correlated random effects allows us to test the overidentification restrictions implicit in conventional fixed-effects models, which are nested within the unrestricted correlated random effects model (see Ashenfelter and Zimmerman 1997). Since the seminal contribution to cross-subject identifications of teacher effects by Dee (2005), available fixed-effects estimators (commonly implemented in first-differenced estimations) implicitly impose that teacher effects are the same across subjects. Rather than initially imposing such a restriction, after estimating the system of equations (3a) and (3b) we can straightforwardly test whether $\beta_1 = \beta_2 = \beta$ and whether $\eta_1 = \eta_2 = \eta$. Only if these overidentification restrictions cannot be rejected, we can also specify correlated random effects models that restrict the β and η coefficients to be the same across subjects:

$$\text{(4a)} \quad y_{i1} = (\beta+\eta)T_{t1} + \eta T_{t2} + (\gamma+\theta_1)U_{t1} + \theta_2 U_{t2} + (\alpha+\chi)Z_i + (\delta+\phi)X_{i1} + \phi X_{i2} + \tau_{t1} + \varepsilon'_{i1}$$

$$\text{(4b)} \quad y_{i2} = \eta T_{t1} + (\beta+\eta)T_{t2} + \theta_1 U_{t1} + (\gamma+\theta_2)U_{t2} + (\alpha+\chi)Z_i + \phi X_{i1} + (\delta+\phi)X_{i2} + \tau_{t2} + \varepsilon'_{i2}$$

This restricted correlated random effects model is equivalent to conventional fixed effects models that perform within-student comparisons in two subjects to eliminate bias from unobserved non-subject-specific student characteristics. They have been applied to estimate effects of non-subject-specific teacher attributes such as gender, ethnicity, and credentials (cf. Dee 2005; 2007; Ammermüller and Dolton 2006; Clotfelter, Ladd, and Vigdor 2007). Note, however, that in order to identify β and γ in specifications like equations (4a) and (4b), it has to be assumed that the assignment of students to teachers is random and, in particular, is not related to students' subject-specific propensity for achievement. Otherwise, omitted teacher characteristics such as pedagogical skills and motivation, comprised in the teacher-specific error component τ_t, could bias estimates of the observed teacher attributes.

In order to avoid such bias from omitted teacher characteristics, we propose to restrict the analysis to samples of students who are taught by the same teacher in the two subjects. In such a setting, $U_{t1} = U_{t2} = U_t$ and $\tau_{t1} = \tau_{t2} = \tau_t$, so that the education production functions simplify to:

$$\text{(5a)} \quad y_{i1} = (\beta+\eta)T_{t1} + \eta T_{t2} + (\gamma+\theta_1+\theta_2)U_t + (\alpha+\chi)Z_i + (\delta+\phi)X_{i1} + \phi X_{i2} + \tau_t + \varepsilon'_{i1}$$

$$\text{(5b)} \quad y_{i2} = \eta T_{t1} + (\beta+\eta)T_{t2} + (\gamma+\theta_1+\theta_2)U_t + (\alpha+\chi)Z_i + \phi X_{i1} + (\delta+\phi)X_{i2} + \tau_t + \varepsilon'_{i1}$$

This model has a straightforward first-differenced representation:

$$y_{i1} - y_{i2} = \beta(T_{i1} - T_{i2}) + \delta(X_{i1} - X_{i2}) + \varepsilon'_{i1} - \varepsilon'_{i2} \quad (6)$$

which is equivalent to including both student and teacher fixed effects in a pooled specification. In this specification, any teacher characteristic that is not subject-specific (U_t and τ_t) drops out.

Identification of teacher effects is still possible in our setting because teacher subject knowledge varies across subjects. Note that this specification can be identified only in samples where students are taught by the same teacher in the two subjects. This makes it impossible to identify teacher effects for attributes that do not vary across the two subjects *within* the same teacher. Thus, this specification cannot identify effects of teacher attributes that are not subject-specific, such as gender and race.[6] But it allows eliminating bias from standard omitted teacher variables when estimating the effect of teacher subject knowledge.

A final remaining concern might be that the matching of students and teachers is not random with respect to their respective *relative* performance in the two subjects. For example, if there are two classrooms in the same grade in a school, in principle it is conceivable that students who are better in math than in reading are assigned to one of the classes and that a teacher who is better in math than in reading is assigned to teach them. Therefore, to avoid bias from any non-random allocation of teachers to students, we finally restrict our main sample to schools that have only one classroom per grade. This eliminates any bias from sorting between classrooms within the grade of a school. While this restriction comes at the cost of identification from a sample of relatively small and mostly rural schools that are not necessarily representative of the total population, additional analyses provided below suggest that the sample restriction is not driving our results. By restricting the analysis to schools in mostly rural regions where parents have access to only one school, the restricted sample additionally eliminates any possible concern of non-random selection of schools by parents based on the relative performance in math versus reading of the schools and the students. Note that, by eliminating non-random teacher selection, the sample restriction also rules out any bias from prior differences in student achievement, as

[6] The within-teacher identification also accounts for possible differences between teachers in their absenteeism from the classroom, an important topic of recent research on teachers in developing countries (cf. Banerjee and Duflo 2006). Note also that in a group of six developing countries with data, teacher absenteeism rates are lowest in Peru (Chaudhury et al. 2006).

students cannot be allocated on grounds of within-student performance differences between the subjects to appropriate teachers.

One way to test that student-teacher assignment is indeed random in our sample with respect to their relative capabilities in the two subjects is to include measures of subject-specific student characteristics (X_{ij}), as well as of teacher characteristics other than subject knowledge that might vary between subjects, in the model. Therefore, in robustness specifications we will include a subject-specific index of student motivation and subject-specific measures of teaching hours and teaching methods in the model. Also, teachers whose first language is not Spanish may not only be relatively better in math than in reading, but may also put relatively less effort into teaching Spanish. Since teachers are likely to be from the same ethnic group as their students, we will test for robustness in a sample restricted to students whose first language is Spanish.

III. Data and Descriptive Statistics

To implement our identification strategy, which is rather demanding in terms of specific samples and testing in two separate subjects of both students and teachers, we employ data from the 2004 Peruvian national evaluation of student achievement, the "Evaluación nacional del rendimiento estudiantil" (EN 2004). The study tested a representative sample of 6th-graders in mathematics and reading, covering a total of over 12,000 students in nearly 900 randomly sampled primary schools. The sample is representative at the national level and for comparisons of urban versus rural areas, public versus private, and multi-grade[7] versus "complete" schools.

The two tested subjects, math and reading, are subjects that are separately taught in the students' curriculum. The student tests were scheduled for 60 minutes each and were developed to be adequate for the curriculum and knowledge of the tested 6th-graders. As a unique feature of EN 2004, not only students but also their teachers were required to take tests in their respective subjects. The teacher tests were developed independently from the student tests, so that the two do not have common items. While in most studies in the existing literature, tested teacher skills can be only weakly tied to the academic knowledge in the subject in which student achievement is examined (cf. Wayne and Youngs 2003), the separate subject-specific teacher tests in EN 2004

[7] Multi-grade schools, in which several grades are served in the same class by the same teacher, are a wide-spread phenomenon in many developing countries where parts of the population live in sparsely populated areas (cf. Hargreaves et al. 2001). For example, the remoteness of communities in the Andes and the Amazon basin means a lack of critical mass of students to provide complete schools that have at least one classroom per grade.

allow for an encompassing measurement of teacher subject knowledge in two specific subjects.[8] Both student and teacher tests in both subjects were scaled using Rasch modeling by the Unit for Quality Measurement (UMC) of the Peruvian Ministry of Education.

Of the 12,165 students in the nationally representative dataset, 56 percent (6,819 students) were taught by the same teacher in math and reading. Of the latter, 63 percent (4,302 students) are in schools that have only one class in 6th grade. The combination of being served by the same teacher in both subjects in schools with only one classroom in 6th grade is quite frequent, as the Peruvian school system is characterized by many small schools spread out over the country to serve the dispersed population. As the main focus of our analyses will be on this "same-teacher one-classroom" (STOC) sample (although we also report OLS estimates for the full sample below), we focus on this sample here.[9]

Table 1 reports descriptive statistics. Both student and teacher test scores are scaled to a mean of 0 and standard deviation of 1. The first language was Spanish for 87 percent of the students, and a native language for the others. As the descriptive test scores in columns (6) and (7) reveal, Spanish-speaking students fare significantly better than students whose first language is not Spanish, although their teachers perform only slightly better. Both students and teachers score better in urban (compared to rural) areas, in private (compared to public) schools, and in complete (compared to multi-grade) schools. In terms of teacher education, 28 percent of the teachers hold a university degree, compared to a degree from a teaching institute. Both math and reading is taught at an average of about 6 hours per week.

Subject-specific inspection (not shown) reveals only a few substantive subject-specific differences in sub-populations. Boys and girls score similar in reading, but boys outperform girls on average in math. While male and female teachers score similarly on the math test, male teachers score worse than female teachers in reading. Teachers with a university degree score worse in math but slightly better in reading compared to teachers with a degree from an institute.

---

[8] Evidence on content-related validity comes from the fact that both teacher test measures do not show a significant departure from unidimensionality. Principal component analysis of the Rasch standardized residuals shows a first eigenvalue of 1.7 in math and 1.5 in reading, indicating that both tests are measuring a dominant dimension – math and reading subject knowledge, respectively. Additional validity stems from the fact that at least five subject experts in math and reading, respectively, certified adequacy of test items for both teacher tests.

[9] While the restricted STOC sample is certainly not a random sample of the full population of Peruvian 6th-graders, a matching procedure reported below indicates that the restriction is not driving our results.

IV. Results

A. Main Results

As a reference point for the existing literature and the subsequent analyses, Table 2 reports OLS regressions of student test scores on teacher test scores and different sets of regressors, pooling test data from the two subjects. Throughout the paper, standard errors are clustered at the classroom level. As the first five columns show, there is a statistically significant association between student test scores and teacher test scores in the full sample, both in a bivariate setting and after adding different sets of controls. The association is substantially reduced, however, when the three school characteristics – urban location, private operation, and being a multi-grade school – are controlled for. Student achievement is significantly positively associated with Spanish as a first language, urban location, private operation, being a complete school, female teacher, and a student motivation index. As the final four columns reveal, the significant association between student and teacher test scores holds also in the sample of students taught by the same teacher in both subjects and in the same-teacher one-classroom (STOC) sample, although point estimates are lower than in the full sample.

The top panel of Table 3 presents reduced-form results of the unrestricted correlated random effects model of equations (3a) and (3b), estimated by seemingly unrelated regressions. By restricting the student sample to those who are taught by the same teacher in both subjects, we avoid bias stemming from within-school sorting. In the same-teacher sample, student math scores are significantly related to teacher math scores after controlling for teacher reading scores. Likewise, student reading scores are significantly related to teacher reading scores after controlling for teacher math scores. While the coefficient estimates on the differing-subject teacher scores are positive in both models, they are smaller than the same-subject coefficients and statistically insignificant. The pattern is the same in the smaller STOC sample, although at a lower level of statistical significance.

The unrestricted model allows us to test for the overidentification restrictions implied by first-differenced models – whether the coefficient on teacher math scores in the math model differs from the coefficient on teacher reading scores in the reading model, and whether the coefficient on teacher reading scores in the math model differs from the coefficient on teacher math scores in the reading model. In both samples, both χ^2 statistics are far from rejecting the hypothesis that the respective coefficients are the same in the two models (see Table 3).

This result allows us to estimate a correlated random effects model that restricts these coefficients to be the same across subjects (bottom panel of Table 3). In this specification, the coefficient on other-subject teacher scores provides an estimate of η in equations (4a) and (4b). As expected, the estimate is positive, albeit only marginally significant, in the same-teacher sample, indicating positive selection effects. The estimate is close to zero in the STOC sample, indicating that selection effects may indeed be eliminated by the sample restriction to students who have the same teacher in both subjects in schools that have only one classroom in the grade. In this specification, the structural parameter of interest, β, which captures the effect of teacher subject knowledge on student achievement, is identified by the difference between the coefficient on same-subject teacher scores and the coefficient on other-subject teacher scores (see equations (4a) and (4b)). As reported in the last row of Table 3, the effect of teacher subject knowledge on student achievement is positive and statistically significant in both samples.

This central result of our paper is replicated in the first-differenced specifications of Table 4. As laid out in equation (6), here the within-teacher within-student model is implemented by differencing test scores across the two subjects (results are the same when implemented in a pooled-subject sample with fixed effects for students and teachers). The first-differenced models confirm a highly significant positive effect of teacher subject knowledge on student achievement. In the same-teacher sample, the coefficient estimate on the teachers' between-subject test-score difference is 0.037, significantly smaller than the OLS estimate reported in Table 1.[10]

In the STOC sample, which excludes any remaining possible bias from sorting of students within the grade, the highly significant point estimate of 0.047 is surprisingly close to the OLS estimate with basic controls in the same sample. This is further support that the teacher-student assignment can indeed be viewed as random in this sample. The magnitude of the coefficient estimates of the within-teacher within-student specification in the two samples is statistically not distinguishable between the two columns. The STOC estimate indicates that an increase in teacher test scores of one standard deviation raises student test scores by about 4.7 percent of a

[10] The slight differences in point estimates between the estimates implied in the restricted correlated random effects models (last row of Table 3) and the estimates of the first-differenced models (Table 4) stem from the fact that the coefficients on the control variables, which are differenced out in the first-differenced model, are allowed to differ across subjects in the correlated random effects model. The first-differenced specification equivalent to this correlated random effects model is provided in column (3) of Table 6 below.

standard deviation. However, as discussed below, this estimate still suffers from attenuation bias due to measurement error in the teacher test scores.

As a test for the randomness of the student-teacher assignment in this sample, Table 5 reports models that add subject-specific measures of student and teacher attributes to the first-differenced within-teacher within-student specification. Results are reported for the STOC sample; results for the larger same-teacher sample, which contains schools with more than one 6th-grade classroom, are qualitatively the same.[11] Column (2) introduces an index of subject-specific student motivation.[12] Subject-specific differences in student motivation may be problematic if they are pre-existing and systematically correlated with the difference in teacher subject knowledge between subjects. On the other hand, they may also be the result of the quality of teaching in the respective subject, so that controlling for differences in student motivation between the two subjects would lead to an underestimation of the full effect of teacher subject knowledge. In any event, although the between-subject difference in student motivation is significantly associated with the between-subject difference in student test scores, its inclusion hardly affects the estimated effect of teacher subject knowledge.

The additional columns of Table 5 introduce subject-specific teacher measures. Because these measures are missing for a considerable number of observations, the basic specification is first reported for the smaller sample and then compared to the model that includes the additional regressors. The point estimate on teacher test scores is slightly lower in the smaller samples (columns (3) and (5)), but statistically indistinguishable from the full STOC sample.

Column (4) introduces the difference in weekly teaching hours between math and reading to control for the possibility that teachers may prefer to teach more in the subject they know better, or get to know the subject better which they teach more. The effect of teaching hours is positive and significant. Increasing teaching time by one weekly hour raises student test scores by about 3 percent of a standard deviation. But the coefficient magnitude of teacher test scores is not affected compared to estimating the basic model on the same sample.

[11] Detailed results are available from the authors on request.

[12] The student motivation index, scaled 0-1, is calculated as a simple average of five survey questions corresponding to each subject such as "I like to read in my free time" (reading) or "I have fun solving mathematical problems" (math), on which students can choose disagree/sometimes true/agree. The resulting answer is coded as 0 for the answer displaying low motivation, 0.5 for medium motivation, and 1 for high motivation.

Column (6) enters controls for differences in how teachers teach the curriculum in the two subjects.[13] Teachers report for each subject whether they use subject-specific books, work books, local, regional, or national guidelines, and other help to design their subject curriculum. Again, the effect of teacher subject knowledge is unaffected by controlling for these other subject-specific teacher characteristics. Together, the models with subject-specific controls confirm the significant causal effect of teacher subject knowledge.

Another concern with the identification, given the relatively large percentage of students for whom Spanish is not the main language, may lie in the fact that teachers who are not native Spanish speakers may not only be relatively worse in teaching Spanish compared to math, but may also exert less effort in teaching Spanish (even after controlling for their knowledge of Spanish). Since teachers are likely to be from the same ethnic group as their students, one way to avoid such bias is to estimate the model on the sample of native Spanish speaking students only. The coefficient on the teacher test score difference in this reduced sample is 0.042 (standard error 0.016), indicating that lower relative effort of non-native Spanish teachers in teaching reading is not driving our results.

The descriptive statistics indicated a few sub-group specific differences between the two subjects. In particular, male students and teachers appear to fare relatively better in math than in reading, compared to the same between-subject difference for females. To ensure that such systematic sub-group differences in between-subject test-score differences are not driving our results, Table 6 adds controls for student, school, and teacher characteristics. Note that all these controls do not vary within students, so that their effect on average test scores is eliminated by the student fixed effects. The specification additionally tests whether the controls are related to the *difference* in achievement between the subjects. While the negative association of relative math achievement with female students and teachers, as well as with public schools, is confirmed in the regression analysis, the estimated effect of teacher subject knowledge is hardly affected.

[13] EN 2004 asks teachers to describe which items they use to design the subject curriculum: working books, school curriculum, institutional educational projects, regional educational projects, 3rd cycle basic curriculum structure, and/or readjusted curricular programs. When using each of them as a differenced dummy in the regression, only one becomes weakly significant and they are jointly insignificant.

B. Correcting for Measurement Error

Any test can only be an imperfect measure of teachers' subject knowledge. As is well known, classical measurement error in the explanatory variable will give rise to downward bias in the estimated effects. Glewwe and Kremer (2006, p. 989) argue that this is likely to be a serious problem in estimating school and teacher effects. Assume that, in any given subject *s*, true teacher test scores T^* are measured with classical measurement error *e*:

$$T_{is}^* = T_{is} + e_{is}, \qquad E(e_{is}) = Cov(T_{is}^*, e_{is}) = 0 \tag{7}$$

Suppose that T^* is the only explanatory variable in an educational production function like equations (1a) and (1b) with mean-zero variables, and e and ε are uncorrelated. Classical measurement error theory tells us that using measured T instead of true T^* as the explanatory variable leads to well-known attenuation bias, where the true effect β is asymptotically attenuated by the reliability ratio λ (e.g., Angrist and Krueger 1999):

$$y_{is} = \beta\lambda T_{is} + \tilde{\varepsilon}_{is}, \qquad \lambda = \frac{Cov(T_{is}^*, T_{is})}{Var(T_{is})} = \frac{Var(T_{is}^*)}{Var(T_{is}^*) + Var(e_{is})} \tag{8}$$

The problem is that in general, we do not know the reliability λ with which a variable is measured.

But in the current case, where the explanatory variable is a test score derived using psychometric modeling, the underlying psychometric test theory in fact provides estimates of the reliability ratio, the most common statistic being Cronbach's α (Cronbach 1951). The method underlying Cronbach's α is based on the idea that reliability can be estimated by splitting the underlying test items in two halves and treating them as separate measures of the underlying concept. Cronbach's α is the mean of all possible split-half coefficients resulting from different splittings of a test and as such estimates reliability as the ratio of true variance to observed variance of a given test. It is routinely calculated as a measure of the reliability of test metrics, and provided as one of the psychometric properties in the EN 2004 data.[14]

[14] In Rasch modeling, there are additional alternative measures of the reliability ratio, such as the person separation reliability. In the EN 2004 teacher test, these are very close to Cronbach's α (at 0.73-0.77 in math and 0.58-0.62 in reading), so that applying them provides very similar (albeit even slightly larger) estimates of the true effect.

Of course, such reliability measures cannot inform us about the test's validity – how good a measure it is of the underlying concept (teacher subject knowledge in our case). But they provide us with information of the measurement error contained within the test metric derived from the underlying test items. Note that differences between the test metric and the underlying concept only reduce total reliability further, rendering our following adjustments still a conservative estimate of the "true" effect of teacher subject knowledge on student achievement.

While measurement error attenuates estimates of any individual test, the measurement-error problem is aggravated even further in our case by the use of first-differenced test-score measures, which additionally reduce the signal-to-noise ratio. Let λ_M be the reliability ratio of the teacher math test and λ_R the reliability ratio of the teacher reading test. Assuming that measurement errors are unrelated across subjects – $Cov(e_M, e_R) = Cov(T_M, e_R) = Cov(e_M, T_R) = 0$ – and using $Var(e_s) = (1-\lambda_s)Var(T_s)$, the bias λ_Δ of the coefficient $\hat{\beta}$ in the first-differenced specification of equation (6) can be derived as:

$$
\begin{aligned}
\lambda_\Delta &= \frac{Var(\Delta T_i^*)}{Var(\Delta T_i^*) + Var(\Delta e_i)} \\
&= \frac{Var(T_M^*) + Var(T_R^*) - 2Cov(T_M^*, T_R^*)}{Var(T_M^*) + Var(T_R^*) - 2Cov(T_M^*, T_R^*) + Var(e_M) + Var(e_R) - 2Cov(e_M, e_R)} \\
&= \frac{\lambda_M Var(T_M) + \lambda_R Var(T_R) - 2Cov(T_M, T_R)}{Var(T_M) + Var(T_R) - 2Cov(T_M, T_R)}
\end{aligned}
\tag{9}
$$

With λ_M and λ_R given by Cronbach's α and $Var(T_M)$, $Var(T_R)$, and $Cov(T_M, T_R)$ observed in the measured test-score data, we can thus calculate the unbiased effect of T^* on y as $\beta = \hat{\beta}/\lambda_\Delta$.

Results are reported in the bottom panel of Table 4. In our main specification, λ_M=0.74, λ_R=0.64, the variance of each test is scaled to 1, and the covariance between the two tests is 0.424 in the STOC sample. This yields an estimate for λ_Δ of 0.462. With a biased estimate $\hat{\beta}$ of 0.047, the true effect of teacher subject knowledge on student achievement turns out to be 0.101: an increase in student achievement by 10.1 percent of a standard deviation for a one standard deviation increase in teacher subject knowledge.

Note that this is an upper bound of the true effect if measurement errors are correlated across subjects. For example, if 10 percent of the observed covariance between the two tests were due to positive covariance between the measurement errors, equation (9) suggests that the true effect would be 0.094 rather than 0.101.

C. Interpretation of Effect Size and Results in Sub-Samples

In order to assess whether this overall effect stems from an exposition of a student with a teacher for one year or for more years, we can observe in the data how long each teacher has been with the tested class. A caveat with such an analysis is that 6th-grade teachers' score differences between math and reading may be correlated with previous teachers' score differences between subjects, although it is not obvious that this a substantial issue in the subject-differenced analysis. It turns out that 39 percent of the STOC sample has been taught by the specific teacher for only one year, an additional 38 percent for two years, and only the remaining 23 percent for more than two years. An interaction term between tenure with class and the teacher test-score difference, although positive, does not capture statistical significance when added to our basic model.

Table 7 reports results of estimating the model separately for these three sub-groups. The point estimate in the sub-sample of students who have been taught by the teacher for only one year is even larger than the point estimate in our main specification, indicating that the estimated effect mostly captures a one-year effect. The coefficient estimate in the sub-sample of teachers that have been with the class for at least three years is even larger. However, the point estimate in the sub-sample of teachers who have been with the class for two years is smaller. While we do not have an explanation for this, we note that statistical power is relatively weak in these smaller samples. The confidence bands, reported in the middle panel of Table 7, suggest that the 2-year-sample estimate may well be larger than the 1-year-sample estimate. (Note that in the larger sample of students with the same teacher in both subjects, but not necessarily only one classroom per grade, the effect in the 2-year sub-sample (2,355 observations) is statistically highly significant, at a point estimate of 0.054.)

Recent evidence suggests that overall teacher effects may tend to fade out relatively quickly, by up to 50 percent per year (Kane and Staiger 2008). The bottom panel of Table 7 reports the expected cumulative effect of teacher subject knowledge in the three sub-samples if the point

estimate of our main specification, 0.047, were a 1-year effect and effects were to fade out by 50 percent per year.[15] It turns out that the expected cumulative effects are well within the confidence bands of the estimates on the three sub-samples. This is still true when assuming smaller fadeouts of 25 percent or 10 percent. As a consequence, while the statistical power of the sub-sample models is limited, we note that they are in line with an interpretation where the estimate of our main specification is viewed as the effect of having been with a teacher of a certain subject knowledge for one year.

In such an interpretation, the measurement-error corrected estimate means that if a student who is taught by a teacher at, e.g., the 5th percentile of the subject knowledge distribution was moved to a teacher with median subject knowledge, the student's achievement would be raised by 16.6 percent of a standard deviation by the end of the school year. If one raised *all* 6th-grade teachers who are currently below the 95th (75th, 50th) percentile on the subject knowledge distribution to that percentile, *average* student performance would increase by 16.2 (8.1, 4.0) percent of a standard deviation. If one raised all teachers in 1st through 6th grade and assumed 25 percent annual fadeout of the teacher effect, average student performance would increase by 53.3 (26.8, 13.2) percent of a standard deviation by the end of 6th grade. Of course, this would mean substantial improvements in teacher knowledge, but ones that appear conceivable when considering the low average teacher performance in the given developing-country context.

It is also illuminating to compare our estimated size of the effect of teacher subject knowledge to the size of the recent estimates of the overall teacher effect based on relative student test score increases exerted by different teachers. Rockoff (2004, pp. 247-248) concludes for the U.S. school system that a "one-standard-deviation increase in teacher quality raises test scores by approximately 0.1 standard deviations in reading and math on nationally standardized distributions of achievement." Rivkin, Hanushek, and Kain (2005) find a similar magnitude. It is impossible to relate our results directly to this finding, both because the Peru finding may not generalize to the United States and because the line-up of teachers along the dimension of student test-score increases will not perfectly match the line-up along the dimension of teacher subject knowledge, so that the two metrics do not necessarily match. But it is still interesting to

[15] The estimate in column (3) is based on the observed share of teachers being with the class for 3, 4, 5, or 6 years, multiplied by the cumulative effect implied after the respective time. Note that with a fadeout of 50 percent per year, the marginal importance of the 4th to 6th year is very limited: the cumulative effect resulting from a one-year-exposition effect of 0.047 is 0.082 after 3 years, 0.088 after 4 years, 0.091 after 5 years, and 0.092 after 6 years.

note that our measurement-error corrected estimate of the effect of teacher subject knowledge falls in the same ballpark.

To test whether our main effect hides effect heterogeneity in sub-populations, we introduce interaction terms between the teacher test-score difference and several indicator variables. The interacted models, reported in Table 8, do not find a single statistically significant difference between any of the following sub-groups: female (compared to male) students, students with Spanish (compared to a native language) as their first language, urban (compared to rural) areas, private (compared to public) schools, complete (compared to multi-grade) schools, male (compared to female) teachers, and teachers with a university (compared to institute) degree. While the point estimates of the interaction terms suggest that the effect may be quantitatively smaller in particular in private schools and for teachers with a university degree, statistical power of the models does not suffice to reject that the effect does not differ between the sub-populations. The difference in teaching hours also does not interact significantly with the teacher test-score difference. In sum, there is little evidence that the effect of teacher subject knowledge differs strongly across these sub-populations.

Although this finding suggests that results can be generalized, the fact that our main sample of analysis – the same-teacher one-classroom (STOC) sample – is not a representative sample of the Peruvian student population raises the question whether results derived on the STOC sample can be generalized to the Peruvian student population at large. Schools with only one 6th-grade class that employ the same teacher to teach both subjects are more likely to be found in rural rather than urban areas and in multi-grade schools, and accordingly on average perform lower on the tests.[16] To test whether the sample specifics prevent generalization, we perform the following analysis. We first draw a 25-percent random sample from the original EN 2004 population, which is smaller than the STOC sample. We then employ nearest-neighbor propensity score matching (caliper 0.01) to pick those observations from the STOC sample that are comparable to the initial population. The variables along which this STOC sub-sample is made comparable to the initial population are student test scores in both subjects, teacher test scores in both subjects, and dummies for different school types (urban, public, and multi-grade). On this sub-sample of the STOC sample that is comparable to the whole population, we estimate our within-teacher

[16] Detailed descriptive statistics for the full sample, as well as detailed results of the following matching analysis, are available from the authors on request.

within-student specification. In a bootstrapping exercise, we repeated this procedure 1,000 times. The regression analyses yield an average teacher test score effect of 0.044, with an average standard error of 0.020, for an average number of 1,904 observations. These results suggest that the teacher test score effect of our main specification is not an artifact of the STOC sample, but is likely to hold at a similar magnitude in the Peruvian 6th-grade student population at large.

Finally, given the apparent gender differences in math vs. reading achievement, and given the recent evidence that being assigned to a same-gender teacher may affect student outcomes (Dee 2005; 2007; Carrell, Page, and West 2010), Table 9 estimates whether the effect of teacher subject knowledge differs significantly between the gender match between students and teachers. When distinguishing all four pairs[17] – female student of female teacher, female student of male teacher, male student of female teacher, and male student of male teacher – the model in column (1) cannot identify statistically significant differences in the size of the effect of teacher subject knowledge, although some of the implied differences in point estimates are substantial. To gain statistical power, the model in column (2) distinguishes only whether the student-teacher match has the same gender or not. This specification finds that the effect of teacher subject knowledge is significantly smaller when a student is taught by a teacher of different gender than when taught by a teacher of the same gender. In fact, the implied point estimate in the sample where gender differs between student and teacher is very small. This suggests that student-teacher gender interactions may be important in facilitating the teaching of subject knowledge.

V. Conclusion

The empirical literature analyzing determinants of student learning has tackled the issue of teacher impact from two sides: measuring the impact of teachers as a whole and measuring the impact of distinct teacher characteristics. Due to recent advances in the first stream of literature using newly available rich panel datasets, it has been firmly established that overall teacher quality is an important determinant of student outcomes, i.e., that teachers differ strongly in their impact on student learning (Hanushek and Rivkin 2010). The second stream of literature examines which specific teacher characteristics may be responsible for these big effects and constitute the unobserved bundle of overall teacher quality. Answers to this question are

[17] The four pairs occur with relatively similar frequency, with each of them making up between 21.3 and 28.4 percent of the population.

particularly useful for educational administration and policy-making. In the empirical examination of teacher attributes such as education, experience, salaries, test scores, and certification, only teacher knowledge measured by test scores has reasonably consistently been found to be associated with student achievement (Hanushek 1986; Hanushek and Rivkin 2006).

However, identifying the causal effects of teacher characteristics on student achievement econometrically is a difficult task. Problems of unobserved student and teacher characteristics and of non-random selection into classrooms are likely to bias the estimates available in the literature. If such omitted variables and selection processes are correlated with the achievement of both teachers and students – as is quite likely in such cases as teacher motivation and pedagogical skills, student effort and ability, parental choice of schools and classrooms, and student placements into classrooms – the available conventional estimates will not capture the true effect of teacher knowledge on student outcomes.

This paper proposed a new way to identify the effect of teacher subject knowledge on student achievement, drawing on the variation in subject matter knowledge of teachers across two subjects and the commensurate achievement across the subjects by their individual students. By restricting the sample to students who are taught by the same teacher in both subjects, and in schools that have only one classroom per grade, this identification approach is able to circumvent the usual bias from omitted student and teacher variables and from non-random selection and placement. Furthermore, possible bias from measurement error in teacher subject knowledge was addressed by reverting to psychometric test statistics on reliability ratios of the underlying teacher tests.

We find a significant effect of teacher subject knowledge on student achievement, drawing on data on math and reading achievement of 6^{th}-grade students and their teachers in Peru. A one-standard-deviation increase in teacher subject knowledge raises student achievement by about 10 percent of a standard deviation. Robustness analyses indicate that the result is unaffected by considering between-subject differences in teaching hours, teaching methods, and student motivation, and by restricting the model to students whose main language is Spanish. If anything, results are larger in the sub-sample of teachers who have been with the class for just one year, suggesting that the evidence is best interpreted as a one-year effect. There is little evidence that the result varies significantly in different sub-populations, such as female or male students, students with Spanish or a native language as their first language, urban or rural areas,

female or male teachers, and teachers with or without a university degree. The only significant heterogeneity indicates that the effect may be larger if students and teachers share the same gender, as opposed to opposite-gender pairs. Overall, the results suggest that teacher subject knowledge is indeed one observable factor that is part of what makes up as-yet unobserved teacher quality.

The results suggest that teacher subject knowledge should be clearly on the agenda of educational administrators and policy-makers. Attention to teacher subject knowledge seems to be in order in hiring policies, teacher training practices, and compensation schemes. However, additional knowledge about the relative cost of improving teacher knowledge – both of different means of improving teacher knowledge and compared to other means of improving student achievement – is needed before policy priorities can be established.

The results have to be interpreted in the given developing-country context with relatively low academic achievement. Although average student achievement in Peru is close to the average of Latin American countries (LLECE 2008), it is far below the average developed country. For example, when Peruvian 15-year-olds took the Programme for International Student Assessment (PISA) test in 2002, their average math performance was a full two standard deviations below the average OECD country, at the very bottom of the 41 (mostly developed) countries that had taken the test by the time (OECD 2003). The extent to which the current results generalize to developed countries thus remains an open question.

References

Aaronson, Daniel, Lisa Barrow, William Sander (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25 (1), pp. 95–135.

Ammermüller, Andreas, Peter Dolton (2006). Pupil-Teacher Gender Interaction Effects on Scholastic Outcomes in England and the USA. ZEW Discussion Paper 06-060. Mannheim: Centre for European Economic Research.

Angrist, Joshua D., Alan B. Krueger (1999). Empirical Strategies in Labor Economics. In Orley Ashenfelter, David Card (eds.), *Handbook of Labor Economics*, Volume 3A, pp. 1277–1366. Amsterdam: North-Holland.

Ashenfelter, Orley, Alan B. Krueger. (1994). Estimates of the Economic Return to Schooling from a New Sample of Twins. *American Economic Review* 84 (5), pp. 1157–1173.

Ashenfelter, Orley, David J. Zimmerman (1997). Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers. *Review of Economics and Statistics* 79 (1), pp. 1–9.

Banerjee, Abhijit, Esther Duflo (2006). Addressing Absence. *Journal of Economic Perspectives* 20 (1), pp. 117–132.

Bedi, Arjun S., Jeffery H. Marshall (2002). Primary School Attendance in Honduras. *Journal of Development Economics* 69 (1), pp. 129–153.

Behrman, Jere R. (2010). Investment in Education: Inputs and Incentives. In Dani Rodrik, Mark Rosenzweig (eds.), *Handbook of Development Economics*, Volume 5, pp. 4883–4975. Amsterdam: North-Holland.

Behrman, Jere R., David Ross, Richard Sabot (2008). Improving Quality versus Increasing the Quantity of Schooling: Estimates of Rates of Return from Rural Pakistan. *Journal of Development Economics* 85 (1-2), pp. 94–104.

Card, David (1999). The Causal Effect of Education on Earnings. In Orley Ashenfelter, David Card (eds.), *Handbook of Labor Economics*, Volume 3A, pp. 1801–1863. Amsterdam: North-Holland.

Carrell, Scott E., Marianne E. Page, James E. West (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. *Quarterly Journal of Economics* 125 (3), forthcoming.

Chamberlain, Gary (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics* 18 (1), pp. 5–46.

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

Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor (2007). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. NBER Working Paper 13617. Cambridge, MA: National Bureau of Economic Research.

Coleman, James S., Ernest Q. Campbell, Carol F. Hobson, James M. McPartland, Alexander M. Mood, Frederic D. Weinfeld, et al. (1966). *Equality of Educational Opportunity.* Washington, D.C.: Government Printing Office.

Cronbach, Lee J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16 (3), pp. 297–334.

Dee, Thomas S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review* 95 (2), pp. 158–165.

Dee, Thomas S. (2007). Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources* 42 (3), pp. 528–554.

Ehrenberg, Ronald G., Dominic J. Brewer (1995). Did Teachers' Verbal Ability and Race Matter in the 1960s? Coleman Revisited. *Economics of Education Review* 14 (1), pp. 1–21.

Eide, Eric, Dan Goldhaber, Dominic Brewer (2004). The Teacher Labour Market and Teacher Quality. *Oxford Review of Economic Policy* 20 (2), pp. 230–244.

Ferguson, Ronald F. (1998). Can Schools Narrow the Black-White Test Score Gap? In Christopher Jencks, Meredith Phillips (eds.), *The Black-White Test Score Gap,* pp. 318–374. Washington, D.C.: Brookings Institution.

Ferguson, Ronald F., Helen F. Ladd (1996). How and Why Money Matters: An Analysis of Alabama Schools. In Helen F. Ladd (ed.), *Holding Schools Accountable: Performance-based Reform in Education,* pp. 265–298. Washington, D.C.: Brookings Institution.

Glewwe, Paul, Michael Kremer (2006). Schools, Teachers, and Education Outcomes in Developing Countries. In Eric A. Hanushek, Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, pp. 945–1017. Amsterdam: North-Holland.

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

Hanushek, Eric A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature* 24 (3), pp. 1141–1177.

Hanushek, Eric A. (1992). The Trade-Off between Child Quantity and Quality. *Journal of Political Economy* 100 (1), pp. 84–117.

Hanushek, Eric A. (1997). Assessing the Effects of School Resources on Student Performance: An Update. *Educational Evaluation and Policy Analysis* 19 (2), pp. 141–164.

Hanushek, Eric A., Steven G. Rivkin (2006). Teacher Quality. In Eric A. Hanushek, Finis Welch (eds.), *Handbook of the Economics of Education*, Volume 2, pp. 1051–1078. Amsterdam: North-Holland.

Hanushek, Eric A., Steven G. Rivkin (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100 (2), forthcoming.

Hanushek, Eric A., Ludger Woessmann (2009). Schooling, Cognitive Skills, and the Latin American Growth Puzzle. NBER Working Paper 15066. Cambridge, MA: National Bureau of Economic Research.

Harbison, Ralph W., Eric A. Hanushek (1992). *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*. A World Bank Book. Oxford: Oxford University Press.

Hargreaves, Eleanore, C. Montero, N. Chau, M. Sibli, T. Thanh (2001). Multigrade Teaching in Peru, Sri Lanka and Vietnam: an Overview. *International Journal of Educational Development* 21 (6), pp. 499–520.

Kane, Thomas J., Douglas O. Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) (2008). *Los Aprendizajes de los Estudiantes de América Latina y el Caribe: Primer Reporte de los Resultados del Segundo Estudio Regional Comparativo y Explicativo*. Santiago, Chile: Oficina Regional de Educación de la UNESCO para América Latina y el Caribe (OREALC/UNESCO).

Luque, Javier (2008). The Quality of Education in Latin America and the Caribbean: The Case of Peru. San Isidro: Abt Associates.

Murnane, Richard J., Barbara R. Phillips (1981). What Do Effective Teachers of Inner-City Children Have in Common? *Social Science Research* 10 (1), pp. 83–100.

Organisation for Economic Co-operation and Development (OECD) (2003). *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*. Paris: OECD.

Rice, Jennifer K. (2003). *Teacher Quality: Understanding the Effectiveness of Teacher Attributes*. Washington, D.C.: Economic Policy Institute.

Rivkin, Steven G., Eric A. Hanushek, John F. Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrica* 73 (2), pp. 417–458.

Rockoff, Jonah E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review* 94 (2), pp. 247–252.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, Douglas O. Staiger (2008). Can You Recognize an Effective Teacher When You Recruit One? NBER Working Paper 14485. Cambridge, MA: National Bureau of Economic Research.

Rothstein, Jesse (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125 (1): 175-214.

Rowan, Brian, Fang-Shen Chiang, Robert J. Miller (1997). Using Research on Employees' Performance to Study the Effects of Teachers on Students' Achievement. *Sociology of Education* 70 (4), pp. 256–285.

Summers, Anita A., Barbara L. Wolfe (1977). Do Schools Make a Difference? *American Economic Review* 67 (4), pp. 639–652.

Tan, Jee-Peng, Julia Lane, Paul Coustère (1997). Putting Inputs to Work in Elementary Schools: What Can Be Done in the Philippines? *Economic Development and Cultural Change* 45 (4), pp. 857–879.

Wayne, Andrew J., Peter Youngs (2003). Teacher Characteristics and Student Achievement Gains: A Review. *Review of Educational Research* 73 (1), pp. 89–122.

World Bank (2007). *Toward High-Quality Education in Peru: Standards, Accountability, and Capacity Building*. A World Bank Country Study. Washington, D.C.: The World Bank.

Table 1: Descriptive Statistics

						Test score if indicator = 1	
	Observations	Mean	Std. Dev.	Min	Max	Student	Teacher
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student test score							
Math	4,302	0	1	-4.089	4.199		
Reading	4,302	0	1	-3.545	3.457		
Difference	4,302	0	0.823	-3.515	3.897		
Teacher test score							
Math	4,302	0	1	-3.037	3.641		
Reading	4,302	0	1	-2.267	2.597		
Difference	4,302	0	1.074	-3.396	3.570		
Student characteristics							
Female	4,302	0.481	0.500	0	1	-0.017	-0.015
1st language Spanish	4,290	0.871	0.336	0	1	0.126	0.031
Motivation index	4,214	0.805	0.152	0.1	1	0.261	0.030
School characteristics							
Urban area	4,302	0.533	0.499	0	1	0.364	0.134
Private school	4,302	0.133	0.340	0	1	0.904	0.293
Complete school	4,302	0.534	0.499	0	1	0.343	0.162
Teacher characteristics							
Female	4,057	0.521	0.500	0	1	0.125	0.071
University degree	4,164	0.284	0.451	0	1	0.115	-0.074
Classroom characteristics							
Teaching hours in the subject	3,932	6.076	1.657	1.5	11.7		
Curriculum design							
Subject-specific books	4,276	0.688	0.417	0	1	0.006	0.043
Student working books	4,276	0.466	0.455	0	1	-0.080	-0.033
Local school guidelines	4,276	0.478	0.469	0	1	0.065	-0.019
Institutional guidelines	4,276	0.258	0.401	0	1	0.202	0.023
Regional guidelines	4,276	0.035	0.167	0	1	-0.186	0.282
National guidelines	4,276	0.736	0.407	0	1	0.054	0.056
Adj. curriculum guidelines	4,276	0.311	0.431	0	1	0.038	0.185
Others	4,276	0.153	0.336	0	1	0.021	-0.020

Sample: Same-teacher one-classroom (STOC).

Table 2: Ordinary Least Squares Regressions

Sample:	Full sample					Same-teacher sample		Same-teacher one-classroom (STOC)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teacher test score	0.241***	0.231***	0.105***	0.095***	0.090***	0.164***	0.069***	0.139***	0.045*
	(0.022)	*(0.020)*	*(0.016)*	*(0.018)*	*(0.018)*	*(0.031)*	*(0.023)*	*(0.042)*	*(0.027)*
Student female		-0.018	-0.018	-0.018	-0.014		-0.014		-0.042
		(0.023)	*(0.019)*	*(0.020)*	*(0.020)*		*(0.026)*		*(0.029)*
Student 1st language Spanish		0.650***	0.739***	0.773***	0.722***		0.739***		0.605***
		(0.100)	*(0.059)*	*(0.068)*	*(0.064)*		*(0.068)*		*(0.068)*
Urban area			0.361***	0.345***	0.336***		0.348***		0.338***
			(0.049)	*(0.056)*	*(0.055)*		*(0.064)*		*(0.072)*
Private school			0.632***	0.685***	0.713***		0.632***		0.642***
			(0.045)	*(0.047)*	*(0.049)*		*(0.086)*		*(0.100)*
Complete school			0.305***	0.299***	0.288***		0.282***		0.379***
			(0.049)	*(0.057)*	*(0.058)*		*(0.071)*		*(0.073)*
Teacher female				0.080**	0.076**		0.048		-0.017
				(0.036)	*(0.035)*		*(0.052)*		*(0.063)*
Teacher university degree				0.009	0.015		0.023		0.062
				(0.037)	*(0.039)*		*(0.057)*		*(0.071)*
Hours				0.006	0.006		0.013		-0.005
				(0.011)	*(0.011)*		*(0.015)*		*(0.018)*
Student motivation					0.472***		0.736***		0.763***
					(0.044)		*(0.060)*		*(0.070)*
Teaching method (8 indicators)					yes		yes		yes
Subject math		0.001	-0.001	0.002	-0.061***		-0.082***		-0.072***
		(0.013)	*(0.011)*	*(0.015)*	*(0.017)*		*(0.019)*		*(0.024)*
Constant	0.000	-0.580***	-1.343***	-1.440***	-1.822***	0.000	-1.869***	0.000	-1.511***
	(0.023)	*(0.101)*	*(0.057)*	*(0.093)*	*(0.114)*	*(0.033)*	*(0.138)*	*(0.045)*	*(0.154)*
F	121.31	53.35	160.33	97.55	63.04	27.15	36.65	11.15	38.94
Observations	24,330	24,286	23,574	17,311	16,289	13,638	10,493	8,604	6,565
Students	12,165	12,143	11,787	10,365	10,056	6,819	5,564	4,302	3,485
Classrooms (clusters)	893	893	867	756	749	521	432	346	284

Dependent variable: student test score. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 3: Correlated Random Effects Models

Sample:	Same-teacher sample		Same-teacher one-classroom (STOC)	
	(1)		(2)	
Unrestricted model:	Math	Reading	Math	Reading
Teacher test score in same subject	0.065**	0.049*	0.047	0.039
	(0.031)	*(0.026)*	*(0.037)*	*(0.034)*
Teacher test score in other subject	0.043	0.015	0.026	-0.017
	(0.029)	*(0.026)*	*(0.036)*	*(0.032)*
χ^2	468.43		353.45	
Observations (students)	6,233		3,936	
Classrooms (clusters)	476		316	
χ^2 (coeff. on same-subject teacher score equal)	0.12		0.02	
Prob > χ^2	*0.731*		*0.898*	
χ^2 (coeff. on other-subject teacher score equal)	0.40		0.59	
Prob > χ^2	*0.528*		*0.444*	
Restricted model:				
Teacher test score in same subject	0.058***		0.044**	
	(0.017)		*(0.020)*	
Teacher test score in other subject	0.028*		0.003	
	(0.016)		*(0.019)*	
χ^2	429.76		341.18	
Observations (students)	6,233		3,936	
Classrooms (clusters)	476		316	
Implied β	0.030**		0.041***	
Prob > χ^2	*0.023*		*0.008*	

Dependent variable: student test score in math and reading, respectively. Regressions in the two subjects estimated by seemingly unrelated regressions (SUR). Regressions include controls for student gender, student 1st language, urban area, private school, complete school, teacher gender, and teacher university degree. Clustered standard errors in the SUR models are estimated by maximum likelihood. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 4: Within-Teacher Within-Student Estimation: First-Differenced (Fixed-Effects) Models

Sample:	Same-teacher sample	Same-teacher one-classroom (STOC)
	(1)	(2)
Regression estimate:		
Difference in teacher test score between math and reading	0.037***	0.047***
	(0.013)	*(0.014)*
F	8.69	10.71
Observations (students)	6,819	4,302
Classrooms (clusters)	521	346
Measurement-error correction:		
Effect of difference in teacher test score, measurement-error corrected	0.082	0.101
λ_{Δ}	0.452	0.462
$Cov(T_M,T_R)$	0.434	0.424
λ_M	0.74	0.74
λ_R	0.64	0.64

Dependent variable: difference in student test score between math and reading. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 5: Introducing Subject-Specific Controls in the Fixed-Effects Model

Sample:	STOC		Dropping observations with missing data on Teaching hours		Dropping observations with missing data on Teaching methods	
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher test score difference	0.047***	0.046***	0.038**	0.040**	0.034**	0.034**
	(0.014)	*(0.014)*	*(0.016)*	*(0.016)*	*(0.016)*	*(0.016)*
Student motivation difference		0.124**		0.124*		0.114*
		(0.062)		*(0.066)*		*(0.068)*
Teaching hours difference				0.028**		0.036***
				(0.013)		*(0.014)*
Teaching methods difference (8 indicators)						yes
Constant	0.000	-0.010	0.008	-0.004	0.006	-0.003
	(0.017)	*(0.018)*	*(0.019)*	*(0.019)*	*(0.019)*	*(0.019)*
F	10.70	6.69	5.73	4.67	4.47	1.52
Observations (students)	4,302	4,113	3,575	3,575	3,431	3,431
Classrooms (clusters)	346	345	297	297	288	288

Dependent variable: difference in student test score between math and reading. Same-teacher one-classroom (STOC) sample. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 6: Controlling for Systematic Between-Subject Differences

	(1)	(2)	(3)
Teacher test score difference	0.045***	0.045***	0.041***
	(0.014)	*(0.015)*	*(0.015)*
Student female	-0.146***	-0.145***	-0.145***
	(0.028)	*(0.028)*	*(0.029)*
Student 1st language Spanish	-0.009	-0.016	-0.003
	(0.046)	*(0.046)*	*(0.046)*
Urban area		-0.056	-0.041
		(0.039)	*(0.039)*
Private school		0.149***	0.153***
		(0.057)	*(0.059)*
Complete school		0.012	0.032
		(0.040)	*(0.043)*
Teacher female			-0.136***
			(0.038)
Teacher university degree			0.010
			(0.038)
Constant	0.079*	0.088*	0.129**
	(0.044)	*(0.047)*	*(0.051)*
F	13.50	7.65	7.67
Observations (students)	4,290	4,290	3,936
Classrooms (clusters)	346	346	316

Dependent variable: difference in student test score between math and reading. Same-teacher one-classroom (STOC) sample. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 7: Results by Tenure with Class

Tenure with class:	1 year	2 years	More than 2 years
	(1)	(2)	(3)
Regression estimate:			
Teacher test score difference	0.070***	0.036	0.100***
	(0.025)	*(0.025)*	*(0.032)*
Constant	-0.046*	0.014	0.089*
	(0.026)	*(0.029)*	*(0.045)*
F	7.80	2.15	9.89
Observations (students)	1,478	1,442	886
Classrooms (clusters)	131	94	72
95% confidence interval:			
Lower bound	0.020	-0.013	0.037
Upper bound	0.119	0.086	0.163
Cumulative effect assuming 0.047 annual effect and different annual fadeouts:			
50% fadeout per year	0.047	0.070	0.086
25% fadeout per year	0.047	0.082	0.126
10% fadeout per year	0.047	0.089	0.162

Dependent variable: difference in student test score between math and reading. Same-teacher one-classroom (STOC) sample. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 8: Effects in Sub-Samples Identified by Interaction Terms

X =	Student female	Student 1st language Spanish	Urban area	Private school	Complete school	Teacher male	Teacher university degree	Teaching hours difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher test score difference	0.042**	0.070*	0.053***	0.052***	0.055**	0.049**	0.064***	0.037**
	(0.018)	*(0.038)*	*(0.019)*	*(0.014)*	*(0.023)*	*(0.020)*	*(0.019)*	*(0.016)*
Interaction of X with teacher test score difference	0.009	-0.028	-0.011	-0.049	-0.014	-0.029	-0.038	-0.006
	(0.024)	*(0.041)*	*(0.028)*	*(0.055)*	*(0.030)*	*(0.030)*	*(0.029)*	*(0.013)*
X	-0.144***	-0.0002	-0.015	0.125**	0.016	0.126***	0.003	0.027**
	(0.028)	*(0.051)*	*(0.034)*	*(0.054)*	*(0.034)*	*(0.036)*	*(0.037)*	*(0.013)*
Constant	0.069***	-0.001	0.008	-0.017	-0.009	-0.056**	-0.003	0.005
	(0.022)	*(0.048)*	*(0.023)*	*(0.017)*	*(0.024)*	*(0.023)*	*(0.021)*	*(0.018)*
F	13.20	3.90	3.90	6.52	3.59	8.59	4.41	4.35
Observations (students)	4,302	4,290	4,302	4,302	4,302	4,057	4,164	3,745
Classrooms (clusters)	346	346	346	346	346	323	338	297

Dependent variable: difference in student test score between math and reading. Same-teacher one-classroom (STOC) sample. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent.

Table 9: Student-Teacher Gender Interactions

	(1)	(2)
Teacher test score difference	0.062**	0.076***
	(0.027)	*(0.020)*
Interaction of [male student and female teacher] with teacher test score difference	-0.025	
	(0.033)	
Interaction of [female student and male teacher] with teacher test score difference	-0.059	
	(0.040)	
Interaction of [male student and male teacher] with teacher test score difference	-0.033	
	(0.040)	
Indicator [male student and female teacher]	0.149***	
	(0.035)	
Indicator [female student and male teacher]	0.134***	
	(0.047)	
Indicator [male student and male teacher]	0.269***	
	(0.048)	
Interaction of [different student-teacher gender] with teacher test score difference		-0.056**
		(0.026)
Indicator [different student-teacher gender]		0.010
		(0.030)
Constant	-0.133***	-0.005
	(0.030)	*(0.024)*
F	8.73	5.02
Observations (students)	4,057	4,057
Classrooms (clusters)	323	323

Dependent variable: difference in student test score between math and reading. Same-teacher one-classroom (STOC) sample. Robust standard errors (adjusted for clustering at classroom level) in parentheses: significance at *** 1, ** 5, * 10 percent