Re-Examining the Role of Teacher Quality In the Educational Production Function

Cory Koedel
University of Missouri

Julian R. Betts*
University of California, San Diego

Revised April 2007

Abstract

This study uses administrative data linking students and teachers at the classroom level to estimate teacher value-added to student test scores. We find that variation in teacher quality is an important contributor to student achievement – more important than has been implied by previous work. This result is attributable, at least in part, to the lack of a ceiling effect in the testing instrument used to measure teacher quality. We also show that teacher qualifications are almost entirely unable to predict value-added. Motivated by this result, we consider whether it is feasible to incorporate value-added into evaluation or merit pay programs.

* The authors thank Andrew Zau and many administrators at San Diego Unified School District, in particular Karen Bachofer and Peter Bell, for helpful conversations and assistance with data issues. We also thank Yixiao Sun, Julie Cullen, Nora Gordon, Mark Appelbaum and participants at the 2006 SOLE conference, particularly our formal discussant Eric Hanushek, for their useful comments and suggestions as well as the Spencer Foundation for research support. The underlying project that provided the data for this study has been funded by the Public Policy Institute of California.

I. Introduction

It has been well established that education plays an important role in determining both economic growth and individual life outcomes (for example, see Katz and Murphy, 1992). This has led to an ongoing interest in the determinants of student achievement, including teacher quality. However, researchers have historically struggled to capture the role of teacher quality in the educational production function. Given the importance of education and the undeniable role played by teachers, how much does variation in teacher quality affect student performance?

The vast majority of the empirical work on teacher quality has relied on observable teacher qualifications to measure teacher quality. As a whole, this body of research suggests that these qualifications are only weakly related to student performance.[1] Therefore, we shift our focus away from teacher qualifications and instead measure teacher quality by value-added to student test scores.[2] Although value-added has been criticized by some, it continues to gain traction among both researchers and policy makers. In fact, proposals to base teacher evaluations on value-added, sometimes involving pay incentives, are becoming increasingly common.[3]

We analyze teacher value-added to student performance on math and reading standardized exams using micro-level data from San Diego elementary schools linking students and teachers at the classroom level. Our results indicate that variation in teacher quality, measured by value-added, is considerably larger than previous research has implied. Our larger variance estimates are

[1] For reviews of this literature, see Hanushek (1986, 1996).
[2] There is a small literature that has shifted its focus to teacher value-added. Recent studies include Rivkin, Hanushek and Kain (2005), Hanushek et al. (2005), Aaronson, Barrow and Sander (2007), Nye, Konstantopoulos and Hedges (2004), McCaffrey et al. (2003), Harris and Sass (2006) and Koedel (2007).
[3] For example, see Gordon, Kane and Staiger (2006). Other examples include non-profit groups like Battelle for Kids in Columbus, OH, which has set up a three-year pilot program that uses value-added as an evaluation tool for teachers in Ohio and the state of Florida.

attributable, at least in part, to the lack of a ceiling effect in the testing instrument that we use to measure teacher quality. Test-score ceilings inhibit students' performance gains as test-score levels rise. These ceilings are quite common in practice and have two important implications for value-added analysis. First, in the presence of a test-score ceiling, estimating teacher effects from a typical value-added specification can lead to an understatement of the variance of teacher quality and, in turn, of the importance of teacher quality as an educational resource. Second, a test-score ceiling will influence individual teachers' value-added estimates. This latter issue is of particular concern if value-added is to be used to evaluate teacher performance.

We relate our value-added measures of teacher quality to the qualifications that primarily determine teacher recruitment, retention and salaries. Our results support previous research indicating that these qualifications are poor predictors of teacher performance. Even upper-bound estimates of the ability of observable teacher qualifications to predict variation in outcome-based teacher quality are very small. Similarly, teachers' salaries are virtually uncorrelated with their value-added to student test scores.

Motivated by the weak link between teacher performance and teacher qualifications, and the growing interest in value-added more generally, we consider the role that value-added estimates might play in determining teacher accountability. When compared to the current standards by which most teachers are judged (observable qualifications), a value-added approach offers a significant improvement in terms of rating teachers based on their contributions to actual student performance. However, in both math and reading, estimation error constitutes a considerable

portion of the individually estimated teacher effects. Therefore, value-added modeling may be better suited as just one component of a more comprehensive system of teacher evaluation.

II. Empirical Strategy

We estimate teacher fixed effects from a value-added model of student achievement in the reduced form:[4]

(1)
$$TestScore_{ijkst} = \alpha_i + TestScore_{ijks(t-1)}\psi + ZipCode_{it}\beta + X_{it}\gamma + Z_{it}\rho + C_{it}\eta + D_{it}^{J(teacher)}\theta + D_{it}^{K(grade)}\pi + D_{it}^{S(school)}\delta + \varepsilon_{it}$$

Equation (1) describes the test-score performance of student *i* taught by teacher *j* in grade *k* at school *s* in year *t*. The model controls for heterogeneity in student ability by including student fixed effects (denoted by α_i).[5] The vectors X_{it}, Z_{it} and C_{it} contain time-varying student-, school- and classroom-level characteristics, respectively. The variables included in these vectors are listed in Table 1. Vectors of indicator variables for teachers, grade levels and schools are also included in the student-achievement specification.

In addition to student fixed effects, our model includes school and zip-code fixed effects. Together, these sets of fixed effects ensure that the model evaluates variation in teacher quality within schools, ignoring any between-school variance. Our methodology is supported by previous empirical work indicating that most of the variation in teacher quality occurs within

[4] Value-added is often modeled in terms of test-score gains. The gainscore specification is a specific case of the general value-added specification in equation (1). We do consider gainscore models in our analysis. As would be expected, teacher-effect estimates from a gainscore model that is analogous to equation (1) are highly correlated with our estimates.

[5] Students and teachers in San Diego are non-randomly assigned to classrooms within schools, highlighting the importance of controlling for student ability in the model of student achievement. In an omitted exercise that is available from the authors upon request, we reject the hypothesis that, within schools and grades, current teachers do not predict previous test-score performance. This result additionally implies that students may be sorted along dimensions that are unobserved.

schools as opposed to between schools (Hanushek et al., 2005; Nye, Konstantopoulos and Hedges, 2004). This is likely to be the case because the degree of sorting of teacher quality across schools, which would drive any between-school variation in teacher quality, is largely dependent on the success of schools in identifying and hiring the best teachers.[6] The empirical evidence suggests that schools may find it very difficult to identify the best teachers and that even if they do, they may choose not to hire them.[7] In our model-sensitivity analysis in Section V, we show that essentially all of the variation in teacher quality in San Diego elementary schools exists within schools.

The potential influence of peer effects is possibly the most worrisome confounding factor in any analysis of teacher quality. To address this issue, our model controls for the year (t-1) achievement of classroom-level peers. We also note that the effect of any systematic ability grouping experienced by students will be largely absorbed at the student level because the student fixed effect will pick up the average peer effect experienced by a given student over the course of the panel. Similarly, we control for class size to prevent variation in class size from being misinterpreted as variation in teacher quality.

As it is written, the model in (1) will produce biased estimates of the coefficients of interest because the demeaned error term will be correlated with the demeaned lagged dependent variable. Therefore, we adopt the method of Anderson and Hsiao (1981) to estimate the

[6] Teachers' preferences for better schools could also affect teacher sorting. However, hiring restrictions imposed by the labor contract between San Diego Unified School District and the teacher's union should substantially limit the effects of teachers' preferences on teacher sorting. This will be discussed in more detail in Section V.

[7] Section VIII of this paper shows that observable teacher qualifications are virtually uncorrelated with outcome-based teacher quality. In addition, numerous studies have documented the weak link between observable teacher qualifications and student performance. See, for example, Aaronson et al. (2007), Angrist and Guryan (2003), Betts (1995), Betts, Zau and Rice (2003), Hanushek (1986, 1996), Kane, Rockoff and Staiger (2006). Also, Ballou (1996) argues that schools may choose not to hire the most qualified teachers even when given the opportunity.

equation. The method involves first-differencing to remove the student fixed effects, and then, to account for correlation between the first-differenced lagged dependent variable and the first-differenced error term, estimating this model using 2SLS, instrumenting for $(TestScore_{ijks(t-1)} - TestScore_{ijks(t-2)})$ with $(TestScore_{ijks(t-2)})$. The key assumption required for this instrumentation to be valid is that the error terms in equation (1) are serially uncorrelated. Although this assumption is not directly verifiable using equation (1), we use the first-differenced error terms to test for serial correlation between the ε_{it}'s and find that this primary assumption is upheld.[8] The first-differenced version of equation (1) is detailed below:

$$\textbf{(2)} \quad (TestScore_{ijkst} - TestScore_{ijks(t-1)}) = (\alpha_i - \alpha_i) + (\widehat{TestScore}_{ijks(t-1)} - TestScore_{ijks(t-2)})\psi$$
$$+(ZipCode_{it} - ZipCode_{i(t-1)})\beta + (X_{it} - X_{i(t-1)})\gamma + (Z_{it} - Z_{i(t-1)})\rho + (C_{it} - C_{i(t-1)})\eta$$
$$+(D_{it}^{J(teacher)} - D_{i(t-1)}^{J(teacher)})\theta + (D_{it}^{K(grade)} - D_{i(t-1)}^{K(grade)})\pi + (D_{it}^{S(school)} - D_{i(t-1)}^{S(school)})\delta + (\varepsilon_{it} - \varepsilon_{i(t-1)})$$

The second term in parentheses on the right-hand side is the fitted value for the test score change from the first stage of the 2SLS procedure.[9] We evaluate the effects of teacher quality on student performance in both math and reading using equation (2).

III. Data

This study is based on panel data from the San Diego Unified School District (SDUSD), following elementary school students and teachers over time. We use student test-score data

[8] The white noise assumption for the error term is verified by evaluating the level of serial correlation between the first-differenced error terms, within students, in the first-differenced version of equation (1) below. The individual ε_{it}'s are serially uncorrelated if the first-differenced error terms are serially correlated with a magnitude of approximately -0.5. For students in which more than one first-differenced equation is estimated, we estimate that the serial correlation between the first-differenced error terms to be -0.47.

[9] Robust standard errors for all 2-stage-least-squares coefficients in this model were generated with one important adjustment. The differenced error term in equation (2) is serially correlated among students with more than one equation in our model. We structurally enforced this property of the error term into the variance-covariance matrix for relevant students.

from the Stanford 9 standardized test for both math and reading from the 1998-99 school year through the 2001-02 school year. Our analysis is based on test-score data from over 16,000 students and we evaluate the effects of over 1,000 elementary school teachers at SDUSD. Students and teachers are linked at the classroom level and an extensive list of school, student and teacher characteristics is available.

The Stanford 9 standardized test is psychometrically scaled such that a one-point gain in student performance at any point in the schooling process is meant to correspond to the same amount of learning. A related characteristic of the Stanford 9 test is that, unlike some other standardized tests, it does not exhibit a pronounced test-score ceiling in math or reading performance (through the 5th grade).[10] This feature of the test makes it a particularly useful tool for measuring the effects of teacher quality on student outcomes throughout the entire range of student achievement as will be discussed in further detail in Section VI.

SDUSD is the second largest school district in California and is quite diverse. The student population is approximately 27 percent white, 37 percent Hispanic, 18 percent Asian/Pacific Islander and 16 percent Black. 28 percent of SDUSD students are English learners, and some 60 percent are eligible for meal assistance. Both of these shares are larger than those of the state of California as a whole. As far as standardized testing performance, students in SDUSD trailed

[10] To check for the presence of a test-score ceiling in our data, we group all students into deciles based on their raw test score level in year (t-2). We then check whether the average test-score gains of students in year (t) are lower for students in higher deciles. In math, there is no relation. However, in reading there is a mild but persistent decline in test score gains as students make progress in the test-score levels distribution. See Appendix F for more details.

very slightly behind national reading averages in 1999-2000. On the contrary, SDUSD students narrowly exceeded national norms in math.[11]

This study focuses on elementary school students because they have the same teacher for the entire day. This removes potentially confounding effects such as teacher spillovers that may be present at the high school level. Because students are tested in 2nd through 5th grade (6th grade is part of middle school at SDUSD), we have up to four years of test scores for each student in the panel. Table 1 details the controls available for students, teachers, classrooms and schools in this study. Appendix A provides additional details about the data used for this project.

Table 1. Description of Key Data Elements

Time-Varying Student Characteristics	Controls for grade levels, parental education, level of test score in year (t-1), EL or non-EL (EL = English Learner), FEP or non-FEP (FEP = Fully English Proficient), student was accelerated a grade, held back a grade, a school changer, terms attended, school days attended, student was re-designated FEP that year, student was new to district.
Time-Varying School Characteristics	Controls for the racial makeup and heterogeneity of schools, school size, whether school is year-round, percent of school on free lunch, percent of school EL, percent of school FEP, number of peer coaches, number of peer coach apprentices, percent of school that changed schools, percent of school new to district
Time-Varying Classroom Characteristics	Class size, peer achievement in year (t-1)
Teacher Characteristics	Dummy variables to control for subject of undergraduate degree, undergraduate minor, whether undergraduate institution is a top 100 university based on research dollars, highest level of education, subject of highest degree, level of credentialing, experience, salary, time at SDUSD, controls for any supplementary authorizations, emergency authorizations, and CLAD (Cross-cultural Language and Academic Development) or Bilingual CLAD certification

[11] District characteristics summarized from Betts et al. (2003).

IV. Results – The Variance of Teacher Quality

In this section we evaluate the importance of variation in teacher quality as a determinant of student performance in math and reading. Table 2 reports Wald statistics generated under the null hypothesis that all teacher effects are equal. Variation in teacher quality is shown, quite convincingly, to be a statistically significant determinant of student achievement for both math and reading in elementary school.

Table 2. Wald Tests for the Statistical Significance of Variation in Teacher Quality

	$H_0: \ \theta_1 = \theta_2 = ... = \theta_J = \bar{\theta}$
Math Achievement	Wald Statistic: 2,636 P-Value: < 0.01
Reading Achievement	Wald Statistic: 2,117 P-Value: < 0.01

Although the results in Table 2 indicate that variation in teacher quality is a statistically significant determinant of student achievement, they do not provide information about *economic* significance. To analyze the economic importance of variation in teacher quality as a determinant of student outcomes, we empirically estimate the magnitude of the variance of teacher quality.[12] This will allow us to evaluate the effects of distributional shifts in teacher quality on student performance. We start by calculating the sample variance of the estimated teacher coefficients:

(3) $$Var(\hat{\theta}) = (\frac{1}{J-1})\sum_{j=1}^{J}[\hat{\theta}_j - (1/J)\sum_{j=1}^{J}(\hat{\theta}_j)]^2$$

[12] We follow the method of Koedel (2007) to estimate the magnitude of the variance of teacher quality.

Each fixed-effect coefficient is comprised of two components - the true signal of teacher quality and estimation error, $\hat{\theta}_j = \theta_j + \lambda_j$. Equation (3) overstates the variance of teacher quality because it includes the variance of the estimation error. We define the estimation-error variance as $Var(\lambda)$ and the variance of the teacher-quality signal, the outcome of interest, as $Var(\theta)$. To separate the estimation-error variance from the variance of the teacher-quality signal, we first assume that $Cov(\theta, \lambda) = 0$.[13] This allows for the total variance of teacher fixed effects to be decomposed as follows:

(4) $$Var(\hat{\theta}) = Var(\theta) + Var(\lambda)$$

Next, we scale the Wald statistic and use it as an estimate of the ratio between the total fixed-effects variance and the error variance:[14]

(5) $$\left(\frac{1}{J-1}\right) * [(\hat{\theta} - \bar{\theta}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \bar{\theta}\ell_J)] \approx Var(\hat{\theta}) / Var(\lambda)$$

In the above formulation, $\hat{\theta}$ is the Jx1 vector of estimated teacher fixed effects, $\bar{\theta}$ is the sample average of the $\hat{\theta}_j$'s, $\hat{V}_J$ is the JxJ portion of the estimated variance matrix corresponding to the teacher effects being tested and ℓ_J is a Jx1 vector of ones.[15] Equation (5) weights the total

[13] This assumption is not directly verifiable because both θ and λ are unobserved. If for some reason the signal and error components of teacher fixed effects were negatively correlated then the results presented here would understate the variance of teacher quality. If the converse were the case, the estimates would be overstated.

[14] In the variance matrix that we use for our Wald statistics we set all covariance terms to zero. This covariance restriction has a negligible effect on our results and allows for a straightforward calculation of the magnitude of the variance of teacher quality. See Appendix B for details.

[15] The variance matrix used in the Wald tests is the diagonal of the full variance-covariance matrix for the relevant set of teacher coefficients. Substituting the full variance-covariance matrix for the variance matrix has virtually no effect on the results.

fixed-effects variance by the estimation error variance on a coefficient-by-coefficient basis. See Appendix B for details.

The magnitude of the variance of the teacher-quality signal can be estimated from equations (4) and (5). For example, if the scaled Wald statistic is estimated to be *A* then the variance of the teacher-quality signal can be estimated by:

(6) $$Var(\theta) = Var(\hat{\theta}) - (Var(\hat{\theta}) / A)$$

To facilitate the interpretation of our results, we convert our estimates of the variance of the teacher-quality signal obtained from equation (6) into units of within-grade standard deviations.[16] For math, we estimate that the effect of a one-standard deviation change in teacher quality on student performance is equivalent to 0.26 average within-grade standard deviations of the test. For reading, we estimate an analogous effect of 0.19 average within-grade standard deviations. These results are detailed in the first column of Table 3.[17]

The second column in Table 3 shows the predicted effects of a one-standard-deviation change in teacher quality expressed as a proportion of average annual test-score gains.[18] In math, the effect

[16] To do this we divide the predicted effect on test scores from having a one-standard-deviation increase in teacher quality by the weighted average (across grades) of the standard deviation of end-of-year scores within each grade. The weights are our sample size in each grade. The resulting ratio provides one estimate of the average impact on student performance of a one-standard deviation move upwards in the teacher quality distribution.

[17] The estimates in Table 3 are presented in average within-grade standard deviations of the test that are calculated using all students at SDUSD who have a test-score record. An alternative would be to use only students in our final sample to calculate the average within-grade standard deviations of the test. Estimated within-grade standard deviations based only on students in our sample will be smaller because students used in our sample are more homogeneous than the entire sample at SDUSD (due to the requirements of the fixed effects specification, see Appendix A for details). We ultimately present our estimates using the within-grade standard deviation estimates from the all-student sample because these estimates are likely to be more comparable to others in the literature.

[18] We weight the gains across grades by the sample size in each grade to obtain a weighted average.

of a one-standard deviation change in teacher quality is equivalent to 0.41 student-years. In reading, we estimate an effect of 0.31 student-years.

Table 3. Estimated Effects of Having a One-Standard-Deviation Above-Average Teacher on Student Performance

	Proportion of Average Within-Grade Standard Deviations	Proportion of Average Annual Test-Score Gains
Math	0.26	0.41
Reading	0.19	0.31

The estimates of the variance of teacher quality presented in Table 3 provide strong evidence of the value of teacher quality as a resource in the educational production function and are considerably larger than previous empirical estimates. For example, our estimate of the effect of a one-standard deviation improvement in teacher quality on student math performance is approximately 67 percent larger than an analogous estimate from Hanushek et al. (2005).[19] In both math and reading, we find that significant gains in student performance can be obtained through improvements in teacher quality.

V. Specification Checks and Sensitivity Analysis

The value-added specification of the student-achievement model that we employ, which includes student fixed effects to control for differences in students' test-score trajectories, is unique in the literature. In this section, we evaluate the model in more detail and consider the sensitivity of our variance estimates to alternative specifications.

[19] The 67 percent figure reported in the text is arrived at by taking the raw-gains-scaled estimates from Hanushek et al. (2005) as reported by the authors and comparing them to our estimates. There is an even greater difference between our estimates and those found in Rockoff (2004) and in Rivkin et al. (2005). At the opposite extreme, when compared to estimates from Nye et al. (2004), who use a residual-variance approach that does not correct for sampling variation, our estimates are somewhat smaller.

Table 4 documents four different value-added specifications for the model of student achievement from which teacher fixed effects can be estimated. The first column shows the full model estimated in equation (2). Columns 2 through 4 show three different restricted models. More detail is added to each specification moving from column 2 to column 4. Wald tests for the completeness of the restricted models against the full model indicate that the restricted models in columns 2 and 3 are underspecified.[20]

For each restricted model in Table 4, the bottom two rows of the table compare the vectors of teacher fixed effects estimated from our full model to the given restricted model by reporting the correlation between the vectors. This exercise is performed for the math and reading specifications.

[20] P-values from Wald tests of the null hypotheses that the coefficients on the omitted variables in the restricted models are zero are less than 0.01 for all omitted variable groups except student fixed effects. We do not run tests for the statistical significance of the student fixed effects because of the computational demands of such tests. Furthermore, the large-N, small-T structure of the panel dataset implies that the results from these tests would be rather uninformative (lacking power). However, student fixed effects have a strong theoretical justification for inclusion in the model. For further discussion, see Harris and Sass (2006). Finally, note that all of our major findings are generally robust to models of student achievement that are not first-differenced (see Table 5). The decision about whether to first-difference the value-added specification seems to be most important in determining teachers' value-added rankings (as indicated by Table 4) and merits additional attention in future research.

Table 4. Estimated Correlation Coefficients Relating Teacher Fixed Effects Estimates from Restricted Models to Estimates from the Full Specification

	(1)	(2)	(3)	(4)
Included Explanatory Variables				
Lagged Test Score	Yes	Yes	Yes	Yes
Grade-Level Fixed Effects	Yes	Yes	Yes	Yes
Student-Level Covariates	Yes	No	Yes	Yes
School- and Classroom-Level Covariates, School and Zip Code Fixed Effects	Yes	No	No	Yes
Student Fixed Effects (First Differenced)	Yes	No	No	No
Correlation Coefficient - Math	1	0.64	0.67	0.74
Correlation Coefficient - Reading	1	0.50	0.53	0.62

Notes: Correlation coefficients compare teacher effects weighted by their standard errors. Column 1 shows our full specification to which the restricted specifications in columns 2 through 5 are compared. Wald tests reject all of the restricted models against the full model we have already reported. In columns 2 through 4, the model was estimated without first-differencing.

Why do estimates of teacher quality change so much when we fail to control for unobserved student heterogeneity? One explanation is that teachers are assigned to groups of students in non-random ways based on unobservable student characteristics.[21] Any model that does not control for this will mistakenly attribute inter-student variation in achievement gains to individual teachers. The strong explanatory power associated with student-specific factors implies that models that do not control for these factors may produce biased estimates.

Another explanation is that moving from the between-school specification to the within-student and within-school specification alters the comparison groups for teachers. If there are significant differences in teacher quality across schools at SDUSD, we may wish to compare teachers between as well as within schools. To evaluate this issue we consider the sensitivity of our variance estimates to alternative specifications, including models that exclude both school and

[21] Students do appear to be assigned to classrooms in non-random ways at SDUSD (for example, see Table 5 or footnote 5).

student fixed effects. Table 5 shows eight different models from which we estimate the variance of teacher quality using the variance decomposition in equation (6).[22] The table indicates that the vast majority of the variation in teacher quality among elementary school teachers at SDUSD occurs within schools.

[22] Beyond evaluating the sensitivity of our variance estimates to alternative specifications, we also consider the possibility that our variance estimates are inflated because class-size reductions in California have increased the number of inexperienced teachers at SDUSD relative to other non-California locales. To do this, we separately estimate the variance of teacher quality among experience groups with more/less than two years, more/less than three years, and more/less than 5 years of experience. In line with our findings in Section VIII of this paper, we find that differences in teacher experience explain just a small portion of the variance of teacher quality. For example, the variance of quality among teachers with a sample-average of three years of experience or less is just 5 percent larger than the variance of teacher quality across the entire sample. Ultimately, our interest is in the total variation in teacher quality experienced by students and because of this we do not control for teacher experience directly in our models.

Table 5. Teacher Fixed Effects Variance Estimates, Adjusted Using Equation (4), from Various Math and Reading Student-Achievement Specifications

	Test-Score Levels				Value-Added			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Explanatory Variables								
Lagged Test Score	No	No	No	No	Yes	Yes	Yes	Yes
Student-Level Covariates	No	Yes	Yes	Yes	No	Yes	Yes	Yes
School- and Classroom-Level Covariates, School and Zip-Code Fixed Effects	No	No	Yes	Yes	No	No	Yes	Yes
Student Fixed Effects	No	No	No	Yes	No	No	No	Yes
Estimated Variance of Teacher Quality – Math Model (Standard Deviation in Parenthesis)	527.7 (23.0)	290.2 (17.0)	259.9 (16.1)	86.3 (9.3)	134.2 (11.6)	114.4 (10.7)	115.7 (10.8)	99.5 (9.8)
Estimated Variance of Teacher Quality – Reading Model (Standard Deviation in Parenthesis)	632.3 (25.1)	293.0 (17.1)	190.5 (13.8)	41.6 (6.5)	67.1 (8.2)	52.3 (7.2)	53.4 (7.3)	62.8 (7.6)

Note: For the specifications that omit student fixed effects, additional time-invariant student-level characteristics are included into the models (specifically, information on race and gender) and errors are clustered at the student level. All models include indicator variables for students' grade levels.

The first vertical panel of Table 5 (columns 1 – 4) evaluates teacher effects estimated from a test-score-levels specification. Changes in the variance estimates moving from left to right in this panel show the importance of the various components of the student-achievement model in removing sorting bias based on test-score levels. The second vertical panel evaluates teacher effects estimated from our value-added specification.

We start by estimating the variance of average test-score levels, conditional on students' current grade levels, across teachers at SDUSD. These estimates are presented in column 1 of the table and incorporate not only teacher quality, but also any sorting of students and teachers throughout SDUSD across schools and classrooms. In moving from column 1 to column 2, we add our set of student-level variables to the test-score-levels specification. The variance estimates fall by approximately 50 percent for both the math and reading models. This indicates that observable student-level variables control for a sizeable portion of the district-wide sorting that is contributing to the variance estimates in column 1. In moving from column 2 to column 3, the inclusion of the set of school- and classroom-level covariates and school and zip-code fixed effects further reduces the estimated variance of teacher quality. One possible explanation for this effect is that test-score-levels sorting bias is reduced. That is, student sorting across schools that is aligned with test-score performance, in levels, is removed by the inclusion of these controls. Another possibility is that variation in teacher quality due to teacher sorting across schools is removed from the total variance estimates. Finally, we add student fixed effects to the levels specification in column 4 to control for any within-school sorting of students and teachers that is not captured by observables. The estimates in column 4 show that there is a significant degree of positive student-teacher matching within schools based on students' test-score levels.

The inclusion of student fixed effects significantly reduces the estimated variance of the conditional teacher means at SDUSD by removing upward bias generated by this matching.

We also estimate the variance of the estimated teacher effects across models within the value-added framework. These results are presented in columns 5 – 8. The pattern of adjustments in the variance of the conditional teacher means when moving across models in the value-added framework is quite similar to the pattern displayed in the levels specifications with two important exceptions. First, in both math and reading, school-level variables do not affect the magnitude of the estimated variance of teacher quality in the value-added framework. This implies that although teachers may sort themselves based on observable student characteristics, there is virtually no sorting of teacher quality across schools at SDUSD conditional on these observable student characteristics. This lends strong support to our empirical approach that estimates teacher value-added within schools and students. Second, in the value-added reading model, the inclusion of student fixed effects into the otherwise fully specified model leads to a very mild *increase* in the estimated variance of teacher quality. Given positive student-teacher matching, we would expect the opposite effect.

Estimates from columns 6 and 7 in Table 5 indicate that there is virtually no between-school variation in teacher quality, measured by value-added, across San Diego elementary schools. The lack of between-school variation in teacher value-added is likely to be largely the result of the inability of schools to identify and hire the best teachers. In Section VIII, we show that the observable teacher qualifications most commonly linked to teacher recruitment, retention and

salaries are almost entirely unable to predict teacher value-added.[23] Furthermore, Ballou (1996) shows that even when schools are able to hire seemingly superior teachers, they often choose not to. Finally, schools at SDUSD are further limited in their ability to select the most effective teachers by the labor contract between SDUSD and the teachers' union. This contract requires that schools with an open position choose from the five teachers with the most district seniority who apply for the position and meet the stated qualifications, restricting each school's pool of potential applicants.[24] Overall, the results from Table 5 suggest that the conventional wisdom that there is significant variation in teacher value-added between schools at the elementary level may be quite inaccurate.[25]

Column 8 of Table 5 shows that the inclusion of student fixed effects in the value-added model of student achievement does not significantly inflate the magnitude of the estimated variance of teacher quality in either subject. In fact, for math, moving to the student-fixed-effects specification results in a decrease in the estimated variance of teacher quality. This is intuitive because this specification reduces the bias generated by positive student-teacher matching within schools. Nonetheless, previous researchers who have estimated outcome-based teacher quality have tended to exclude student fixed effects from the value-added specification, presumably because of a belief that the student-fixed-effects model artificially inflates the estimated variance of teacher quality by adding noise to the model of student achievement. A comparison of our math and reading results in Table 5 provides insight into this concern. We find that the student-

[23] For additional evidence, see Aaronson et al. (2007), Angrist and Guryan (2003), Betts (1995), Betts et al. (2003), Hanushek (1986, 1996) and Kane et al. (2006).

[24] Empirical evidence suggests that experience beyond the first few years of teaching is, at most, marginally related to teacher value-added.

[25] This conventional wisdom is likely borne from differences in observable teacher qualifications across schools that are easily documented. However, the link between these observable teacher qualifications and actual teacher value-added is so weak that differences across schools along this dimension provide no information about differences across schools in terms of actual teacher quality as measured by value-added.

fixed-effects specification can lead to inflated variance estimates (for example, mildly in our reading specification), but that this apparently counterintuitive effect is easily explainable. In both math and reading, controls for student ability remove omitted variables bias in teacher fixed effects generated by positive student-teacher matching. However, in our reading analysis, properties of the testing instrument used to measure teacher quality are such that the bias created by this matching is *downward*. The next section explores this issue in detail.

VI. Estimating the Variance of Teacher Quality and the Testing Instrument

The use of the Stanford 9 standardized exam at SDUSD is a fortuitous circumstance for our evaluation of teacher quality. Unlike other testing instruments that have recently been used to estimate outcome-based teacher quality, the Stanford 9 exam is not a minimum competency test. Minimum competency tests are likely to exhibit strong ceiling effects characterized by students experiencing systematic declines in test-score gains as they advance in the test-score levels distribution.[26] Importantly, a test-score ceiling may affect more than just the highest achievers. Appendix F details the test-score ceiling properties of the Stanford 9 standardized exam at SDUSD and shows that the math portion of the Stanford 9 does not exhibit a test-score ceiling at all. For reading, the Stanford 9 exhibits a mild test-score ceiling.

Test-score ceilings are a major consideration in the estimation of outcome-based teacher quality because they restrict the capacity of the testing instrument to capture the full extent of students' human capital development. Hanushek et al. (2005) report that in their analysis of one large Texas school district, gains in test scores are strongly negatively related to previous performance.

[26] Such a relationship will exist for any testing instrument due to regression to the mean. However, in addition to any effects from regression to the mean, minimum competency tests should exert additional downward pressure on test-score gains as students make progress in the test-score levels distribution.

They show that approximately two-thirds of the students in their sample (those at the top of the test-score levels distribution) are at a level of achievement such that the average annual test-score gain of students in their same achievement-level decile is *negative*.[27] Rockoff (2004) does not examine test-score ceiling effects in his analysis in great detail, but does indicate that 3 to 6 percent of students in his study have test scores that are at the maximum attainable score.[28] Other studies fail to address this important issue altogether.

To illustrate how a test-score ceiling can affect estimates of the variance of outcome-based teacher quality, consider a simple example. Teacher effects are estimated using the value-added framework, but suppose that the modeling strategy does not control for unobserved student ability in gains. Assume, as is the norm, that students and teachers are positively matched in terms of ability within schools and that the most able students tend to have larger test-score gains and therefore, higher test-score levels. First, consider a testing instrument given to students that does not exhibit a test-score ceiling. That is, the average gain for high-achieving students is not structurally restricted to be lower than the average gain for low-achieving students by the test. In the absence of controls for student ability, positive student-teacher matching in this scenario will result in a bias away from zero for all teacher fixed effect estimates.[29] This is because the best teachers will be matched with the brightest students (those with the highest gains) and the worst teachers with the students for whom gains are most difficult. This will inflate the estimated variance of teacher quality.

[27] The strength of the negative relationship reported by these authors implies that ceiling effects, in addition to any regression to the mean, are a relevant concern in their analysis.
[28] For comparison, just 0.09 and 0.077 percent of students in our math and reading samples respectively scored at the top score possible for their grade.
[29] Assuming that the distribution of teacher effects is centered around zero. More generally, the bias will be away from the center of the teacher-effect distribution, increasing the variance.

Second, consider the same scenario of positive student-teacher matching in terms of ability but instead imagine a testing instrument that exhibits a test-score ceiling. In this case, lower-performing students will be able to achieve higher test-score gains, on average, simply because of the structure of the test (an example of such a test would be a minimum competency test). Again, the best teachers will teach the most able students but instead of generating an upward bias in teacher effects, these teachers will instead be penalized by the test because their students' gains will be suppressed. Similarly, the worst teachers will be rewarded by the test because their students' gains will be, relatively speaking, overstated. In this scenario, the variance of teacher quality will be understated because both the best and worst teachers will have coefficient estimates that will be biased *toward* zero as a result of positive student-teacher matching.

Now consider the inclusion of controls for student ability in the model of student achievement in both of the above scenarios. In the first scenario, where there is not a test-score ceiling, the inclusion of student fixed effects will remove the upward bias in the teacher fixed effects and reduce the estimated variance of teacher quality. This effect can be seen in moving from column 7 to column 8 in Table 5 for the math analysis, where we find no evidence of a test-score ceiling at SDUSD (see Appendix F). In the second scenario, where a test-score ceiling is present, the inclusion of student fixed effects will again remove bias associated with positive student-teacher matching. However, we will observe the opposite effect on the estimated variance of teacher quality because positive student-teacher matching creates bias *toward* zero in the teacher fixed effects. The inclusion of student fixed effects removes this bias and the estimated variance of teacher quality actually *increases*. This effect can be seen in moving from column 7 to column 8

in Table 5 for the reading analysis, where we find evidence of a test-score ceiling at SDUSD (see Appendix F).[30] Although the effect of the inclusion of student fixed effects on the estimated variance of teacher quality works in opposite directions in these different scenarios, it removes bias from the same source in both cases – positive student-teacher matching. Finally, note that the ceiling effects in our reading analysis are quite mild. In a minimum competency testing environment, a test-score ceiling could have an effect that is significantly more pronounced.

VII. Correlation of Teacher Effectiveness Across Subjects: Math & Reading

Using the teacher coefficients estimated from the models of student achievement for math and reading, we examine the correlation of teacher quality across subjects. Because elementary school students typically stay with the same teacher for the entire day, this question is of particular relevance for this study.

We estimate the correlation coefficient between $\hat{\underline{\theta}}_m$ and $\hat{\underline{\theta}}_r$ (the vectors of teacher coefficients estimated from the math and reading specifications, respectively) to be 0.35. However, this correlation defines the relationship between ($\underline{\theta}_m + \underline{\lambda}_m$) and ($\underline{\theta}_r + \underline{\lambda}_r$), not $\underline{\theta}_m$ and $\underline{\theta}_r$ (where $\underline{\lambda}_m$ and $\underline{\lambda}_r$ represent estimation error). Furthermore, the relationship between $\underline{\lambda}_m$ and $\underline{\lambda}_r$ is unclear *a priori*. Following Rockoff (2004), if we assume that the correlation of true teacher quality across subjects for all teachers is the same, we can get an idea of the direction of the bias introduced by the measurement error in the estimated teacher fixed effects. Measurement error will be smaller for teachers with a greater number of student-year observations. Therefore, we

[30] Relative to other studies, the test-score ceiling present in the reading analysis here is very weak, which in turn explains why its effect on our variance estimates is small. However, the very fact that the estimated variance of teacher quality, measured in terms of reading performance, does not decline when student fixed effects are added to the value-added model is an indication of the ceiling effect.

compare the correlation coefficient between $\underline{\hat{\theta}}_m$ and $\underline{\hat{\theta}}_r$ for a subset of teachers who have a relatively high number of students to that of the entire teacher sample to get an idea of the direction of the effect of the correlation between $\underline{\lambda}_m$ and $\underline{\lambda}_r$ on our initial correlation estimate. The estimated correlation coefficient from our selected subset of teachers is higher than its counterpart from the full teacher set. Thus, measurement error is biasing our estimate of the correlation of teacher quality across subjects toward zero.[31] We present our estimate of the correlation between $\underline{\hat{\theta}}_m$ and $\underline{\hat{\theta}}_r$, 0.35, as a lower-bound estimate of the correlation of teacher quality across subjects.

To estimate an upper bound on the correlation of teacher quality across subjects, we estimate the correlation between $\underline{\theta}_m$ and $\underline{\theta}_r$ under the assumption that the correlation between $\underline{\lambda}_m$ and $\underline{\lambda}_r$ is zero (See Appendix C for details). Our upper-bound estimate of the correlation coefficient relating teacher quality across subjects is 0.64. Overall, our bounded estimate (0.35 to 0.64) indicates that the ability to be an effective teacher, at least at the elementary level, does not appear to be strongly subject-specific.

VIII. Teacher Fixed Effects and Observable Teacher Qualifications

Because variation in outcome-based teacher quality has been shown to be such an important contributor to student achievement, it is of interest to identify observable teacher qualifications that are strong predictors of teacher performance. We use a second-stage regression to evaluate the ability of a rich set of observable teacher qualifications to predict teacher value-added as

[31] Our finding in this regard is in accordance with Rockoff (2004).

estimated by our empirical model. Many of the observable qualifications used in this analysis are important determinants of teacher recruitment, retention and salaries.

The SDUSD dataset includes over 50 unique observable teacher qualifications that may predict teacher value-added. However, running the "kitchen sink" model yields limited information due to collinearity among these qualifications. Therefore, we initially include only key observable qualifications that are unlikely to be highly collinear in our model. We report results using both the smaller model and the model containing all of the observable teacher qualifications available in the dataset (for a listing of the controls used in the richer model, see Table 1).

Consider the following second-stage regression that we would like to estimate:

(7) $$\theta_j = \alpha + X_j\beta + e_j$$

Here, θ_j is the true measure of teacher quality for teacher j in either subject, X_j is a vector of observable teacher qualifications, α is an intercept and e_j is the unobserved error term. However, in the second stage, our dependent variable is a statistical estimate and thus is measured with error.

(8) $$\hat{\theta}_j = \theta_j + \lambda_j$$

The estimation error, λ_j, will appear in the second-stage error term. We would like to estimate α and β from equation (7) above. However, because of the estimation error in the dependent variable, we must estimate the following equation:

(9) $$\hat{\theta}_j = \alpha + X_j\beta + \lambda_j + e_j$$

Here, λ_j and e_j are assumed to be uncorrelated and λ_j may be non-symmetric. The appropriate estimation strategy for efficient estimates of α and β under these circumstances is WLS. The appropriate variance-covariance matrix to use for weighting, following Borjas and Sueyoshi (1994), is:

$$\Omega = \hat{\sigma}_e^2 I_J + \hat{V}$$

where J is the number of teacher coefficients and $\hat{V}$ is a diagonal matrix whose elements are from the diagonal of the estimated variance-covariance matrix corresponding to the teacher coefficients from equation (2). $\hat{V}$ estimates the variance matrix of λ_j. $\hat{\sigma}_e^2$ can be estimated following Borjas (1987). Table 6 reports our FGLS coefficient estimates from the weighted regression.[32,33]

[32] Regressors for our second-stage analysis are averaged within teachers where relevant.

[33] Despite empirical evidence indicating that teacher experience is non-linearly related to effectiveness, we model it linearly here. This is because the linear experience term maximizes the R^2 from the OLS analog to the GLS model presented in the text. (It maximizes the GLS R^2 as well, although the GLS R^2 is difficult to interpret). In an auxiliary analysis available from the authors upon request, we also estimate our second-stage model using experience indicator variables rather than the linear term. Our results from that analysis are virtually identical to those presented in the text.

Table 6. Dependent Variables: Estimated Teacher Coefficients from Equation (2) in Section II for Math and Reading

Variable	Math Analysis	Reading Analysis
Teacher Experience	0.29*	0.21
	(0.13)	(0.12)
School Top 100	-0.98	0.04
	(1.19)	(1.04)
Full Credential	4.80	-1.98
	(2.94)	(2.55)
Master's Degree	0.18	0.60
	(0.97)	(0.84)
BA Education	1.78	0.40
	(0.99)	(0.86)
BA Social Science	3.27*	0.46
	(1.11)	(0.96)
BA English	-1.67	0.16
	(1.83)	(1.59)
BA Math	-3.90	-3.00
	(7.39)	(6.41)
Math Supplemental Authorization	7.35*	4.21
	(3.69)	(3.14)
Art Supplemental Authorization	2.18	4.12
	(3.21)	(2.78)
Language Supplemental	-0.01	3.63
Authorization	(2.81)	(2.43)
R^2	0.0341	0.0138
Adj. R^2	0.0198	-0.0007

* Significant at 5% level of confidence.
Standard errors in parentheses.
Observable teacher qualifications are averaged over time within teachers where relevant.
Teacher experience has been capped at 10 years. That is, teachers with over 10 years of experience are input as having 10 years of experience. It is a well-established fact that the returns to teaching experience decline significantly as experience increases. Indeed, if teaching experience were not capped at 10 years, then experience would cease to significantly predict effective teachers.
The variable 'School Top 100' indicates whether the undergraduate institution attended by the teacher was in the top 100 universities in terms of research dollars.
Supplementary authorizations are obtained by completing a required set of college courses in the field of the authorization. These authorizations are not required for any elementary school teachers.

Rather than focusing on causality, we instead consider the overall power of observable teacher qualifications to predict variation in outcome-based teacher quality. Although the FGLS estimates presented in Table 6 are efficient given the estimation error in the teacher fixed effects, R^2 statistics generated from GLS models have an unclear interpretation (for example, these statistics are not bounded on the interval [0,1]). Therefore, to provide an in-depth answer to the

question of how much variation in teacher quality can be explained by observable teacher qualifications, we use the R^2 formula from the OLS analogs to the above GLS models.

Following the methodology outlined in Appendix D, we generate upper bounds on the R^2 statistics for our math and reading second-stage models by manually removing the variation in the dependent variable due to estimation error from the explanatory-power calculation. These upper bounds estimate the absolute maximum amount of information about variation in actual teacher quality contained by easily observable teacher qualifications. For math, we estimate an upper bound on the true R^2 from our second-stage analysis of approximately 0.057. For reading, the estimated upper bound is just 0.029. Even these upper bounds clearly show that observable teacher qualifications are weak predictors of variation in outcome-based teacher quality.

We also consider an expanded version of our second-stage model that includes all of the observable teacher qualifications available in the data (see Table 1).[34] In this case, we estimate upper bounds of 0.070 and 0.068 for the math and reading analyses respectively. However, we note that our upper bound results are more likely to be overstated with this larger model. See Appendix D for details.

Finally, we consider the unlikely scenario that schools are already identifying effective teachers in ways that evade our methodology and that this identification is reflected in teacher salaries. We run another second-stage regression to see how well teacher salaries alone predict teacher

[34] This expanded model includes indicator variables for undergraduate minors, credential levels, CLAD and BCLAD ((Bilingual) Cross-Cultural Language and Development) certifications, additional supplementary authorizations, additional undergraduate majors and additional advanced degrees. We also include a separate variable that controls for experience at SDUSD specifically.

quality to test for this possibility. We generate *upper bounds* on the percentage of variation in teacher quality explained by teacher salaries to be just 1.4 percent in math and 0.9 percent in reading. This result suggests that teacher compensation, which in SDUSD as in most public school districts depends heavily on teacher tenure, highest degree and teaching credentials, bears almost no relation whatsoever to teaching effectiveness.

IX. Teacher Fixed Effects and Teacher Evaluations

The weak link between outcome-based teacher quality and the qualifications by which most teachers are evaluated should perhaps encourage the use of alternative measures of quality. Among educational-accountability advocates, one proposal is to incorporate output from models similar to our own into teacher evaluations directly (for example, see Gordon, Kane and Staiger, 2004).[35]

To assess the feasibility of using statistically estimated teacher coefficients for teacher evaluations, we first examine whether they contain a sufficiently large signal of actual teacher quality. For math, our variance decomposition in Section IV indicates that the variance of the teacher-quality signal is roughly 60 percent of the total fixed-effects variance. For reading, 50 percent of total fixed-effects variance represents the true signal of quality. Because the relative magnitudes of the signal and noise components of the individual teacher coefficients will be reflective of the entire sample, on average, we use these distribution-wide estimates as estimates

[35] An initial concern is whether teachers should be evaluated within or between schools. Because Table 5 shows that virtually all of the variation in teacher value-added at SDUSD occurs within schools and that there is a considerable degree of within-school student sorting, we use the full within-school and within-student specification documented in equation (2) in our teacher-evaluation analysis. We consider the costs associated with this strategy in Tables 8 and 9 below.

of the average signal-to-noise ratios that characterize the individually estimated teacher fixed effects in math and reading.

On the one hand, these estimates indicate that the teacher-quality signal contained by the value-added coefficients represents a significant improvement over current methods, as discussed in the previous section. However, the high levels of estimation error inherent in the individual fixed effects make their application to teacher evaluation or merit pay programs worthy of a cautious approach.

To illustrate the potential consequences associated with the noise found in our estimates we examine the persistence of estimated teacher fixed effects across years. For this analysis, we focus on student math performance.[36] We break our student sample into two separate subsets based on the year of the differenced dependent variable from equation (2) in Section II. For the first group, the dependent variable in equation (2) is the difference between spring 2002 and spring 2001 test scores. For the second, the dependent variable is the difference between spring 2001 and spring 2000 test scores. We reference the first group as "year t" and the second group as "year t-1". After separating our sample, we independently estimate equation (2) and generate two separate vectors of teacher coefficients, one from each subset of student data. The teacher coefficients estimated from these data subsets are based on different but partially overlapping groups of students. We evaluate the effects of the 941 teachers (out of our initial sample of 1,064) who taught students in both subsets.

[36] Dividing our student sample into two distinct student subsets and performing our analysis separately for each of these subsets introduces substantial noise into our teacher coefficient estimates. In our math analysis, teacher coefficient estimates retained enough signal to make the split-sample analysis possible. However, in reading the estimation error introduced by splitting our sample increased the estimation error variance so much that informative analysis was not possible because the signal-to-noise ratio was close to zero.

Following a methodology similar to that of Aaronson, Barrow and Sander (2007), we examine the rank-persistence of teacher fixed effects across the student subsets. Within each vector of teacher fixed effects we divide teachers into quintiles based on their value-added rankings where quintile-5 teachers are those with the highest value-added. Table 7 demonstrates the persistence of these quintile rankings across the data subsets.

Table 7. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages)

		Teacher Coefficient Quintile Ranking From Year t				
		1	2	3	4	5 (best)
Teacher Coefficient Quintile Ranking From Year t-1	1	**30**	20	19	18	13
	2	23	25	13	21	18
	3	18	20	25	24	13
	4	15	16	26	20	23
	5 (best)	13	17	16	19	**35**

Note: (N = 941). Teachers are placed into quintiles using coefficient estimates from each data subset separately, quintile 5 being the best. Rows sum to 100 percent.

If teacher quality were perfectly observable through statistical estimation and constant over time, entries along the diagonal of Table 7 would all equal 100 percent and all off-diagonal entries would all equal 0. Clearly, this is not the case. In fact, significant fractions of teachers move up or down by two quintiles or more when we shift our student sample.[37] However, the southeast and northwest corners of Table 7 suggest that the best and worst teachers (who are ranked in the top and bottom quintiles) are significantly more likely to retain their distinctions across years relative to other teachers in the sample. Although this result is largely by design (these quintiles are open-ended), it is nonetheless an important feature of this analysis because it is precisely

[37] Importantly, the coefficients evaluated in Table 7 contain much higher levels of estimation error than their counterparts from our full model. This is the result of splitting our student sample because, in doing so, we reduce the number of observations available to estimate each teacher coefficient. The increased estimation error will lead to an understatement of the persistence of teacher effects. An additional concern is that the length of our panel forces us to overlap two of the four years of student data to perform the split-sample analysis. Through this overlap, the correlation between the two sets of teacher fixed effects may be artificially *increased* because the errors in the two sets of estimates may be positively correlated.

these teachers who would be targeted by a teacher-accountability system. Therefore, the bleak outlook portrayed in Table 7 may be somewhat mitigated when considered in the context of an evaluation system focusing on the identification the best and worst teachers.

One concern in our split-sample analysis is that it will understate the persistence of teacher effects as a result of our within-school-and-student specification. This is because the stability estimates from the transition matrix in Table 7 are affected by changes in teachers' comparison groups as teachers move in and out of schools over time. Although teacher movement over time would affect even a between-school analysis, its effects are amplified by our within-school-and-student approach because each teacher's comparison group is smaller and therefore more responsive to teacher turnover.[38]

We present two additional transition matrices analogous to the one in Table 7 to evaluate this concern. The first matrix is generated from a between-school-and-student specification (this specification omits school-level covariates and school- and student-level fixed effects, see column 6 in Table 5) and is detailed in Table 8. The second is still based on the within-school-and-student specification but only uses data from a given school if the average teacher taught at that school in at least three out of the four years of the data panel (84 out of the 108 elementary schools used in this analysis were designated as "low-turnover" by this standard). This matrix is detailed in Table 9.

[38] Another concern here could be that teachers' quality levels may be changing over time with experience. Although the results from Section XIII indicate that experience is only weakly related to value-added, we nonetheless look to see if more experienced teachers have more stable value-added estimates. If experience plays a non-negligible role, we should expect relatively inexperienced teachers to have less stable value-added coefficients because performance has been shown to change most rapidly in the early years of teachers' careers. We do not find any evidence that more experienced teachers have more stable value-added estimates.

Table 8. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages) – Between-Schools-and-Students Specification

		Teacher Coefficient Quintile Ranking From Year t				
		1	2	3	4	5 (best)
Teacher Coefficient Quintile Ranking From Year t-1	1	43	29	14	10	4
	2	26	21	25	18	9
	3	12	21	28	25	15
	4	10	19	19	28	23
	5 (best)	8	11	11	19	50

Note: (N = 941). Teachers are placed into quintiles using coefficient estimates from each data subset separately, quintile 5 being the best. Rows sum to 100 percent.

Table 9. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages) – Within-Schools-and-Students Specification, Low-Turnover Schools Only

		Teacher Coefficient Quintile Ranking From Year t				
		1	2	3	4	5 (best)
Teacher Coefficient Quintile Ranking From Year t-1	1	35	25	16	14	11
	2	19	27	23	15	15
	3	18	20	20	25	17
	4	14	21	18	23	25
	5 (best)	12	9	25	24	29

Note: (N = 824). Teachers are placed into quintiles using coefficient estimates from each data subset separately, quintile 5 being the best. Rows sum to 100 percent.

The tight comparisons among teachers created by our within-school-and-student specification do appear to affect the persistence of teacher effects across the student subsets. In Table 8 the contrast is most stark; looking between schools results in a large increase in the persistence of teacher effects and significantly reduces the percentage of teachers who move more than one quintile in either direction in the transition matrix. Of course, this increased persistence reflects not only the more stable comparison group for each teacher (all teachers in the district rather than just the teachers at a given school) but also the persistence of school-level effects that are correlated with teacher effects.

The differences between Table 7 and Table 9, where we look at low-turnover schools, are more subtle. Although the sums of the diagonal elements of each matrix are very similar, there are significant reductions in the number of teachers who move more than one and more than two quintiles across the transition matrix when we focus our analysis on schools with lower teacher turnover.

Together, the transition matrices in Tables 8 and 9 show that teacher turnover can play an important role in determining year-by-year teacher fixed effects estimated using the within-school-and-student specification. This implies that year-by-year value-added estimates may represent an infeasible standard for evaluating teacher quality.

X. Conclusion

We show that teachers vary in quality considerably more than previous research has implied. In math, we find that the average effect on student performance of a one-standard deviation improvement in teacher quality in a given year corresponds to 0.26 average within-grade standard deviations in test scores. In reading, the same improvement in teacher quality corresponds to 0.19 average within-grade standard deviations. These are very large effects.

Our analysis highlights the importance of the testing instrument used to evaluate teacher quality. We show that when a test-score ceiling restricts students' test-score gains, teacher effects can be significantly understated. However, including controls for heterogeneity in student test-score growth (i.e., student fixed effects) in the value-added specification may at least partially mitigate this problem.

Given the importance of variation in outcome-based teacher quality as a determinant of student achievement, we test to see if the qualifications by which most teachers are evaluated are related to actual performance measured by student outcomes. Our empirical results strongly support earlier findings that observable teacher qualifications are only weakly related to outcome-based measures of teacher quality. To emphasize this, we estimate upper bounds on the explanatory power of observable teacher qualifications and show that even at these bounds, the information about teacher quality contained by these observable measures is minimal. The persistence of this result throughout the modern empirical literature should perhaps lead to long-term changes in teacher recruitment, as well as teacher credentialing and professional development. Perhaps most of all, the system for setting teacher pay largely as a function of teacher experience, education and credentials may require radical reform. We show that teachers' salaries can explain, at most, 0.9 to 1.4 percent of actual variation in performance-based teacher quality.

Finally, the future role of value-added as a determinant of teacher accountability is still unclear. On the one hand, the signal contained by value-added estimates is sizeable, especially when compared to the current standards by which most teachers are evaluated. However, on the other, there is also a considerable degree of estimation error in the teacher coefficients which suggests a cautious approach to their implementation for accountability purposes. One solution would be to incorporate value-added into a larger system of teacher accountability. Employing value-added estimates in conjunction with other measures of teacher quality that are unlikely to have correlated measurement errors should diminish the impact of these errors and increase the visibility of actual teacher quality.

Appendix A
Data Appendix

Section II illustrates the statistical model that seems most appropriate for accurately describing student test-score performance. Specifically, the model accounts for numerous sources of variation in student achievement including variation due to student fixed effects, all within the value-added framework. The structure of the model excludes the use of some of the SDUSD data in that it requires at least three contiguous test scores per student for full identification. However, we require this data restriction in order to specify the most accurate statistical model of student performance possible. Because our entire analysis hinges on the soundness of our teacher fixed effects estimates, the importance of a properly specified model of student performance from which teacher fixed effects are estimated cannot be overstated. Table A1 details the differences between the final sample of students used in our analysis and the general elementary student population at SDUSD.

Table A1. Key Differences Between the Entire SDUSD Elementary Student Sample and the Final Sample Used for Estimation

	All Students	Students with 3 + Years of Data
Race		
% White	26%	28%
% Black	16%	14%
% Asian	17%	20%
% Hispanic	40%	38%
% English Learners	21%	14%
SAT 9 Math Score*	0	0.18
SAT 9 Reading Score*	0	0.20
Avg. Percentage of School on Free Lunch	63%	59%

Our final sample includes 16,303 unique students with at least 3 student-years of data out of a possible 29,973 students who would have been eligible to be included in our model based on the year that they started the 3rd grade.
*Test score performance is measured in average standard deviations from the "All Students" mean (by grade).

As would be predicted, our analysis is based on students who appear to be slightly advantaged relative to the SDUSD population as a whole. However, our final student sample is still reasonably diverse and generally representative of the student population at SDUSD. The biggest difference between the two student populations is in terms of testing performance. Note that the "all students" sample includes students who are movers in the sense that they do not have three contiguous test scores. Thus, Table A.1 is consistent with the well-documented negative relationship between student mobility and performance (see, for example, Rumberger and Larson, 1998; or Ingersoll, Scamman and Eckerling, 1989).

With respect to teachers, we must also be careful about inclusion in our model. Kane and Staiger (2002) find strong evidence of the significant impact of sampling variation on the outcomes of incentive systems based on school-level mean performance measures in North Carolina. Particularly, they find that schools with the smallest populations are considerably more likely to

receive rewards or sanctions based on student performance because the variance of the average of students' test scores from year to year is highest in small schools. A magnified version of this problem arises in our teacher analysis.

By virtue of the general structure of elementary education, elementary school teachers teach just a small number of students each year. Even in studies such as this where numerous years of data are available for each teacher, there are still relatively few data points with which to estimate teacher fixed effects. Particularly in cases where class sizes fluctuate significantly across teachers, or drop to extremely low levels more generally, the impact of sampling variation can dwarf any true signal. Therefore, in an effort to reduce this inherent noise, we restrict our teacher sample to teachers with at least 20 student-years of data. This threshold was chosen as it corresponds to approximately one year of teaching a full elementary classroom. The mean elementary class size in our full dataset is 22.5 students with a standard deviation of approximately 5.5. Thus, a teacher with the mean number of students in her classroom can afford to have up to two students dropped for one reason or another and still be used in our study. Furthermore, this standard removes many teachers who have taught particularly few students. The mean number of student-years per teacher among the dropped teachers was approximately eight. The selection of different student-year cutoff points from as low as 17 student-years to as high as 30 student-years of data reveal no significant changes in our general results beyond the expected mild changes in the precision of teacher coefficient estimates.

Again, restricting our sample of teachers restricts the population for which our results are relevant. Table A2 details key differences between the entire SDUSD elementary teacher population and the sample used in this study.

Table A2. Key Differences Between the Entire SDUSD Elementary Teacher Sample and the Final Sample Used for Estimation

	All Elementary Teachers	Teachers in Our Final Sample
Years Experience	11.08	12.60
% Fully Credentialed	94%	98%
% With Masters Degree	47%	54%
BA Major:		
Education	44%	39%
English	5%	6%
Social Science	21%	26%
Math/Science	2%	2%

Our final sample includes 1,064 teachers from a total of 1,560 potentially eligible teachers available for this study. We define a potentially eligible teacher as a teacher who teaches at least 15 students with at least a current and a lagged test score over the course of the panel. This eligibility requirement would seem to be an absolute minimum for any value-added study. Recall that for our study we require teachers to teach at least 20 students with at least 3 test scores over the course of our panel.
It is often presumed that majors in education are somewhat easier to obtain than majors in other fields (For example, see Ballou, 1996).

With respect to teachers, there is a surprisingly small difference between teachers used in our sample and the entire SDUSD elementary teacher population. Our sample still includes significant variability among teachers in key observable qualifications. After removing teachers with fewer than 20 student-years of data, the average number of student-years of data per teacher in our sample is 37.5**.**

Appendix B
Variance Decomposition

Because the weighting matrix that we use for the Wald statistic is diagonal:

$$(\hat{\theta} - \bar{\theta}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \bar{\theta}\ell_J) = \frac{(\hat{\theta}_1 - \bar{\theta})^2}{\hat{\sigma}_1^2} + \frac{(\hat{\theta}_2 - \bar{\theta})^2}{\hat{\sigma}_2^2} + \ldots + \frac{(\hat{\theta}_J - \bar{\theta})^2}{\hat{\sigma}_J^2}$$

Thus, scaling this summation by the number of teachers returns an estimate of the average ratio of the total fixed-effects variance to the total error variance weighted on a coefficient-by-coefficient basis.

Appendix C
Estimating an Upper Bound on the Correlation of Teacher Value-Added Across Subjects

We generate an upper bound on the correlation of teacher quality across subjects, $corr(\theta_m, \theta_r)$, under the assumption that the correlation coefficient reported in Section VII is understated because $corr(\lambda_m, \lambda_r) = 0$ and this is suppressing our estimate of $corr(\hat{\theta}_m, \hat{\theta}_r)$. Consider the following:

$$corr(\hat{\theta}_m, \hat{\theta}_r) = \{\text{cov}(\theta_m + \lambda_m, \theta_r + \lambda_r) / \{\sqrt{\text{var}(\theta_m + \lambda_m)} * \sqrt{\text{var}(\theta_r + \lambda_r)}\} \quad \text{(C.1)}$$

The correlation coefficient of interest in this analysis is $corr(\theta_m, \theta_r)$. To obtain an upper-bound estimate, we will assume that $\text{cov}(\theta_m, \lambda_r) = 0$, $\text{cov}(\theta_r, \lambda_m) = 0$, and $\text{cov}(\lambda_m, \lambda_r) = 0$ (these conditions also imply that $\text{cov}(\theta_m, \lambda_m) = 0$ and $\text{cov}(\theta_r, \lambda_r) = 0$ because we know that $\text{cov}(\theta_m, \theta_r) \neq 0$) and expect that none of these covariance terms would be negative.[39] Given these conditions we can rewrite equation (C.1) as:

$$corr(\hat{\theta}_m, \hat{\theta}_r) = \{\text{cov}(\theta_m, \theta_r) / \{\sqrt{\text{var}(\theta_m + \lambda_m)} * \sqrt{\text{var}(\theta_r + \lambda_r)}\} \quad \text{(C.2)}$$

By definition, our correlation coefficient of interest is defined as:

$$corr(\theta_m, \theta_r) = \text{cov}(\theta_m, \theta_r) / \{\sqrt{\text{var}(\theta_m)} * \sqrt{\text{var}(\theta_r)}\} \quad \text{(C.3)}$$

Combining C.2 and C.3, we can write:

$$corr(\theta_m, \theta_r) = corr(\hat{\theta}_m, \hat{\theta}_r) * (\sqrt{\text{var}(\theta_m + \lambda_m) / \text{var}(\theta_m)}) * (\sqrt{\text{var}(\theta_r + \lambda_r) / \text{var}(\theta_r)}) \quad \text{(C.4)}$$

This can once again be re-written as:

[39] It is the non-negativity assumption that insures that we are generating an upper bound by setting the covariance of the estimation errors to zero. We justify this assumption by noting that although it is conceivable that there would be a positive correlation between estimation errors for the same classrooms but different subjects, it would be hard to imagine a scenario in which these estimation errors would be negatively correlated.

$$corr(\theta_m, \theta_r) = corr(\hat{\theta}_m, \hat{\theta}_r) * (\sqrt{\sigma^2_{m,fe} / \sigma^2_{m,true}}) * (\sqrt{\sigma^2_{r,fe} / \sigma^2_{r,true}}) \quad \text{(C.5)}$$

Here, $\sigma^2_{-,fe}$ represents the total variance of teacher fixed effects and $\sigma^2_{-,true}$ represents the variance of teacher quality by subject as indicated. We can plug in values for the above variance components using estimates from Section IV. This generates an upper bound estimate of the correlation of teacher effectiveness across subjects of approximately 0.64.

Appendix D
Upper Bound Estimates of the Percentage of Teacher Value-Added Predicted by Observable Teacher Qualifications

The R^2 statistics reported in Table 6 in Section VIII are meant to represent the amount of variation in the teacher coefficients explained by observable teacher qualifications. However, these R^2 values are potentially inaccurate due to measurement error in our second-stage dependent variable and because they are generated from a GLS regression. Our analysis in the text proceeds under the assumption that the measurement error found in our teacher fixed effects coefficients is uncorrelated with observable teacher qualifications.[40] If this is the case, basic R^2 estimates from our second stage analysis will understate the ability of our models to explain true teacher quality because the R^2 statistics are implicitly allowing for the models to predict the measurement error in the dependent variable (which they do not do by assumption). In this appendix, we establish upper bound estimates of the R^2 statistics from our second-stage regressions under the assumption that observable teacher qualifications do not predict the measurement error in our teacher coefficients. If this assumption is incorrect, results from this appendix will over-state the predictive power of observable teacher qualifications.

The GLS estimation performed in Section VIII of the text is used to generate efficient estimates of our coefficients of interest. However, because R^2 statistics from GLS models are difficult to interpret, we proceed here with R^2 statistics from the OLS analogs to the models described in the paper. In order to generate an upper bound on the percentage of variation in true teacher quality

[40] Beyond being very plausible, this assumption is also useful for generating upper bound estimates of the R^2 statistics from our second-stage models. If observable teacher qualifications were somehow predicting the measurement error in the teacher fixed effects even slightly, estimates presented in this appendix will be overstated.

explained by observable characteristics, first consider the general R^2 formula that is estimated by standard software packages for our second-stage analysis:

$$R^2 = 1 - (SSE/SST) \quad \text{(D.1)}$$

$$= 1 - [\sum_{j=1}^{J} (y_j - \hat{y}_j)^2] / [\sum_{j=1}^{J} (y_j - \bar{y})^2] \quad \text{(D.2)}$$

The R^2 formula in equation (D.2) is a consistent estimate of:

$$1 - [E(y_j - \hat{y}_j)^2] / [E(y_j - \bar{y})^2] \quad \text{(D.3)}$$

In this equation, the y_j's correspond to the estimated teacher fixed effects coefficients from the first stage, the $\hat{y}_j's$ are the fitted values of the estimated teacher coefficients from our OLS second-stage regression, and $\bar{y}$ is the mean of the first-stage estimated teacher coefficients. The y_j's can be decomposed as follows:

$$y_j = y_{jtrue} + \lambda_j \quad \text{(D.4)}$$

Here, y_{jtrue} represents true teacher quality and λ_j represents the contribution of estimation error. Substituting equation (D.4) into equation (D.3) yields:

$$1 - [E(y_{jtrue} + \lambda_j - \hat{y}_j)^2] / [E(y_{jtrue} + \lambda_j - \bar{y})^2] \quad \text{(D.5)}$$

Because y_{jtrue} and λ_j are uncorrelated by assumption, the denominator of the second term in equation (D.5) simplifies to $[Var(y_{jtrue}) + Var(\lambda_j)]$. With regard to the numerator, we will continue under the prior that the predictive power of observable teacher qualifications is being understated because observable teacher qualifications do not predict the estimation error in our dependent variable. Therefore, in the spirit of estimating an upper bound we can assume that $\hat{y}_j$ and λ_j are also uncorrelated. Equation (D.5) can be written as:

$$1 - [E(y_{jtrue} - \hat{y}_j)^2 + Var(\lambda_j)] / [Var(y_{jtrue}) + Var(\lambda_j)] \quad \text{(D.6)}$$

If observable teacher qualifications do not predict the estimation error, the above formula adds a positive number representing the variance of the estimation error into both the numerator and denominator of the second term as shown in equation (D.6). Because this term is subtracted from one, this results in an unequivocal understatement of the R^2 reported from our second-stage model.

We can remove the variance of the estimation error from both the numerator and denominator of the second term to estimate an upper bound on the true level of explanatory power exhibited by observable teacher qualifications:

$$1 - [E(y_{jtrue} - \hat{y}_j)^2] / [E(y_{jtrue} - \bar{y}_{true})^2] \quad \text{(D.7)}$$

Using our empirical results from Section IV and the $\hat{y}_j's$ from our second stage regression, we estimate equation (D.7) with:

$$R^2 = 1 - [\sum_{n=1}^{N} (y_{jtrue} - \hat{y}_j)^2] / [\sum_{n=1}^{N} (y_{jtrue} - \bar{y}_{true})^2] \quad \text{(D.8)}$$

It is clear to see how any incidental correlation between the $\hat{y}_j's$ and the λ_j's will lead to an overstatement of this statistic, and thus it is presented as an upper bound. As reported in the text, our upper bound estimates on the explanatory power of observable teacher qualifications are 0.057 and 0.029 for math and reading respectively.

Appendix E
Teacher Quality and Different Student Types

To provide a test of whether teacher effectiveness varies by initial student achievement, we split our student records into two groups based on initial student achievement. Specifically, for each student record, we compare the student's year (t-2) test score to the grade-level median test score for their grade.[41] The first group consists of students who performed at or above the median level of achievement in year (t-2), the second of students who performed below the median. We assign an indicator variable equal to 1 if a student record belongs to the first group and 0 otherwise.

Next, we interact this achievement indicator variable with each of our teacher indicator variables.[42] We then add this new set of interaction terms to the full specification outlined in Section II. The interaction terms will pick up any differences in teacher quality experienced by high-achieving students relative to low-achieving students. That is, if teachers affect different student types differently on a per-teacher basis, then we should find that the set of interaction terms are jointly significant in explaining variation in student performance. However, we find no evidence that the impact of teacher quality varies by student type. For both math and reading, Wald tests fail to reject the null hypothesis that the coefficients on all of the interaction terms are zero. For both math and reading, the p-values from these Wald tests are greater than 0.9.

[41] For example, if a student was in third grade in year (t-2), we look to see if his or her test score in third grade was above or below the third-grade median test score in our sample.

[42] A small percentage (less than 2% for each subject) of the teachers in our sample had all of their students in one achievement group or the other. For these teachers, their interaction terms were dropped from the model.

Appendix F
Test-Score Ceiling Properties at SDUSD

The Stanford 9 standardized test used at SDUSD does not exhibit a test-score ceiling in math and exhibits only a mild-test score ceiling in reading through the 5th grade. As discussed in Section VI of the text, this feature of the Stanford 9 makes it a better instrument with which to measure the variance of teacher quality than some tests used in previous studies. In this appendix, we detail the test-score ceiling properties of the Stanford 9 for both math and reading.

Earlier work with the dataset revealed evidence of some regression to the mean in test scores. This makes it difficult to test for pure ceiling effects by plotting test-score gains in period (t) vs. test score levels in period (t-1) because in part there should be a negative relationship between the two because of regression to the mean. Therefore, to test for the presence of a test-score ceiling in our data, we group all students into achievement deciles based on their raw test score level in period (t-2). We then look to see if the average test-score gains of students in period (t) are lower for students in higher deciles. Figures F.1 and F.2 describe our findings. For math, the Stanford 9 standardized test does not appear to exhibit a test score ceiling. For reading, there is a mild but persistent decline in student test-score gains as students move up in the period (t-2) test-score levels distribution.[43]

[43] Hanushek et al. (2005) present a figure similar to figure F.1 in their analysis. However, in their study, students are grouped into achievement deciles based on period (t-1) test scores, thus combining any test-score ceiling effects with regression to the mean. If we replicate our figures in this appendix following their methodology, we observe a negative relationship for both math and reading as would be expected due to regression to the mean. However, the magnitude of the decline in average test score gains is significantly less in our data when we replicate their analysis and average test-score gains are positive for all student-achievement deciles.

Figure F.1
Achievement Gains by Decile: Math

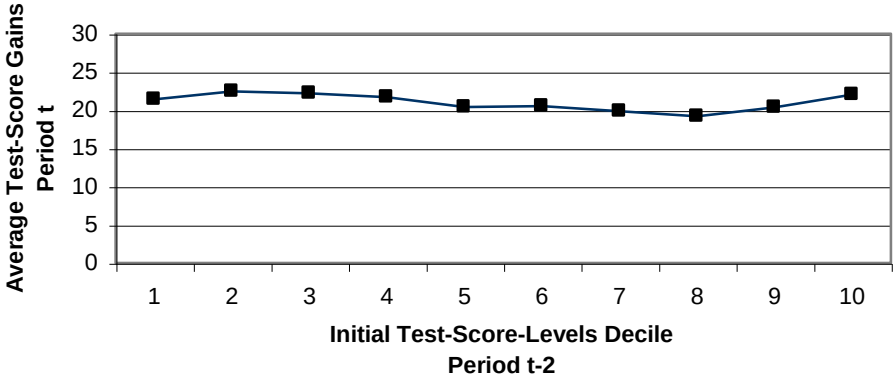
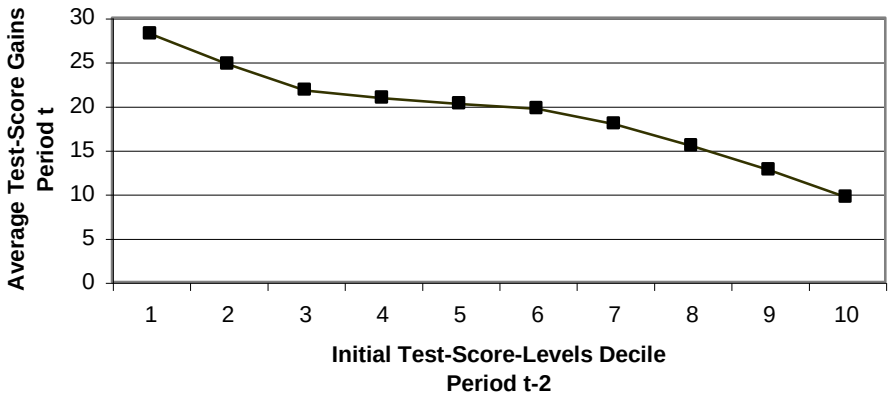


Figure F.2
Achievement Gains by Decile: Reading



References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25:95-135.

Anderson T.W. and Cheng Hsiao. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18:47-82.

Anderson T.W. and Cheng Hsiao. 1981. Estimation of dynamic models with error components. *Journal of American Statistical Association* 76:598-606.

Angrist, Joshua and Jonathan Guryan. 2003. Does teacher testing raise teacher quality? Evidence from state certification requirements. Working Paper no. 9545, National Bureau of Economic Research, Cambridge, MA.

Ballou, Dale 1996. Do public schools hire the best applicants. *Quarterly Journal of Economics* 111:97-133.

Betts, Julian, Andrew Zau, and Lorien Rice. 2003. *Determinants of student achievement, new evidence from San Diego*. Public Policy Institute of California.

Betts, Julian R. 1995. Does school quality matter? Evidence from the national longitudinal survey of youth," *The Review of Economics and Statistics* 77:231-250.

Borjas, George and Glenn Sueyoshi. 1994. A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64:165-182.

Borjas, George. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77:531-553.

Gordon, Robert, Thomas J. Kane and Douglas Staiger. 2004. *Identifying effective teachers using performance on the job.* The Brookings Institution.

Hanushek, Eric, John Kain, Daniel O'Brien and Steven Rivkin. 2005. The market for teacher quality. Working Paper no. 11154, National Bureau of Economic Research, Cambridge, MA.

Hanushek, Eric. 1996. Measuring investment in education. *The Journal of Economic Perspectives* 10:9-30.

Hanushek, Eric. 1986. The economics of schooling: production and efficiency in public schools. *Journal of Economic Literature* 24:1141-77.

Harris, Douglas and Tim R. Sass. 2006. Value-added models and the measurement of teacher quality. Unpublished manuscript, Department of Economics, Florida State University, Tallahassee.

Ingersoll, Gary M., James P. Scamman and Wayne D. Eckerling. 1989. Geographic mobility and student achievement in an urban setting. *Educational Evaluation and Policy Analysis* 11:143-149.

Kane, Thomas E., Jonah E. Rockoff and Douglas O. Staiger. 2006. What does certification tell us about teacher effectiveness? Evidence from New York City. Working Paper no. 12155, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas and Douglas Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16:91-114.

Katz, Lawrence and K. Murphy. 1992. Changes in relative wages, 1963-1987: supply and demand factors. *Quarterly Journal of Economics* 107:35-78.

Koedel, Cory. 2007. Teacher quality and educational production in secondary school. Working Paper, University of Missouri, Columbia.

McCaffrey, Daniel, J.R. Lockwood, Daniel M. Koretz and Laura S. Hamilton. 2003. *Evaluating value-added models for teacher accountability*. RAND Corporation.
Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26:237-257.
Rivkin, Steven, Eric Hanushek and John Kain. 2005. Teachers, schools and academic achievement. *Econometrica* 79:417-458.
Rockoff, Jonah. 2004. The impact of individual teachers on student achievement: evidence from panel data. *American Economic Review*, Papers and Proceedings.
Rumberger, Russell W. and Katherine A. Larson. 1998. Student mobility and the increased risk of high school dropout. *American Journal of Education* 107:1-35.