UNIVERSITY OF CALIFORNIA, SAN DIEGO

Three Essays on Teacher Quality and Educational Production

A Dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Economics

by

Cory Robert Koedel

Committee in Charge

Professor Julian Betts, Chair
Professor Mark Appelbaum
Professor Julie Cullen
Professor Bud Mehan
Professor Yixiao Sun

2007

This Dissertation of Cory Robert Koedel is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2007

To my parents, whose continued support has made my successes possible…

TABLE OF CONTENTS

LIST OF FIGURES

APPENDIX FIGURES

LIST OF TABLES

APPENDIX TABLES

ACKNOWLEDGEMENT

I would like to thank Professor Julian Betts, the chair of my committee. I also thank my other committee members, Julie Cullen, Yixiao Sun, Bud Mehan and Mark Appelbaum. Finally, I would also like to acknowledge Nora Gordon and Kate Antonovics for their assistance with my research and the Spencer Foundation for research support.

Chapter 1, in part, has been submitted for publication as it appears to the Review of Economics and Statistics. The dissertation author was the sole author of this paper.

Chapter 2, in part, has been submitted for publication as it appears to the Journal of Labor Economics with Julian Betts. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is being prepared for publication. The dissertation author was the sole author of this paper.

VITA

2000	Bachelor of Arts, Economics, University of California, San Diego
2004	Master of Arts, Economics, University of California, San Diego
2007	Doctor of Philosophy, Economics, University of California, San Diego

PUBLICATIONS

Julian R. Betts, Lorien Rice, Andrew Zau, Y. Emily Tang, and Cory R. Koedel, *Does School Choice Work? Effects on Student Integration and Achievement,* San Francisco, Public Policy Institute of California, 2006.

FIELDS OF STUDY

Major Field: Economics

Studies in Applied Economics
Professors Julian Betts, Julie Cullen, Yixiao Sun

ABSTRACT OF THE DISSERTATION

Three Essays on Teacher Quality and Educational Production

by

Cory Robert Koedel

Doctor of Philosophy in Economics

University of California, San Diego, 2007

Professor Julian Betts, Chair

This dissertation consists of three essays on teacher quality and educational production. As opposed to the larger body of research on teacher quality, which measures quality primarily by observable qualifications, these essays measure teacher quality in terms of student outcomes. Chapters 1 and 2 of the dissertation measure the effects of teacher quality on test-score outcomes, focusing on teacher value-added, and chapter 3 examines graduation outcomes. The dissertation addresses both elementary- and secondary-school teachers. One theme common to all three papers is that although teacher qualifications are only weakly related to student performance, strong differences in teacher quality emerge when quality is measured in terms of student outputs. That is, student performance can be significantly affected by

distributional shifts in teacher quality. In all three chapters, considerable attention is paid to the methodology behind estimating teacher effects based on student outcomes. Finally, this research also considers the practical application of outcome-based measures of teacher quality for teacher evaluation or accountability.

Chapter 1
Teacher Quality and Educational Production in Secondary School

This study uses administrative data linking students and teachers at the classroom level to evaluate teacher quality and joint production in secondary school. Teacher quality is measured by value-added to student test scores in math and reading. Although empirical research has struggled to link observable teacher qualifications to student achievement, teacher quality measured by student performance varies significantly and has important effects on educational outcomes. I identify which teacher inputs affect which test-score outputs in secondary school and find strong evidence of joint production. The results from this study are applicable to incentive design and teacher accountability.

I would like to thank Andrew Zau and many administrators at San Diego Unified School District, in particular Karen Bachofer and Peter Bell, for assistance with data issues. I also thank Julian Betts, Julie Cullen, Yixiao Sun, Nora Gordon, and participants at the UCSD applied lunch seminars for useful comments and suggestions and 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 and directed by Julian Betts.

I.I. Introduction

In the 2002-2003 school year alone, nearly 388 billion dollars was spent on U.S. elementary and secondary education with 238 billion dollars going to teacher salaries.[1] Despite this large expenditure afforded to provide teachers, there is relatively little research available that quantifies the extent to which variation in meaningful (outcome-based) measures of teacher quality determine student performance. Furthermore, within the relatively small body of literature that does address this issue, there is an even greater dearth of studies that focus on educational output in secondary school.[2]

Secondary-school educational production is quite different than elementary production, specifically with respect to teachers. While elementary students typically have just one teacher per year, secondary-school students are taught by multiple teachers each year. As would be the case in any joint-production setting, standard concerns associated with the assignment of productivity to individual inputs are relevant. For example, there has not been any research to identify which teacher types (i.e., math, English, science, etc.) affect which educational outputs in secondary school. Similarly, the extent to which teachers across subjects are complements or substitutes in the production function is also unknown.

---

[1] From "Revenues and Expenditures for Public Elementary and Secondary Education: School Year 2002–2003" by Frank Johnson and Jason Hill, U.S. Department of Education.

[2] There is only one published study on outcome-based teacher quality in secondary school - Aaronson, Barrow and Sander (2007). Their study focuses on ninth grade math test scores and (predominantly on) math teachers in Chicago public schools.

These unanswered questions are of particular importance in the context of educational accountability. Because we lack empirical evidence linking teacher quality in the various secondary-school subjects to student performance, and because we do not know the relative importance of teacher quality across subjects insofar as it determines achievement, we would be proceeding blindly if we attempted to create performance-based incentives for teachers. For example, teacher quality in some subjects may spill over into student performance in others. These spillover effects should perhaps be incorporated into teacher evaluations, but only for relevant teacher-subject matches and only if they are properly weighted. Otherwise, teachers' incentives would be poorly aligned with performance and free-riding opportunities could be enhanced.

I measure teacher quality by value-added to student test scores in math and reading. In each tested subject, I estimate teacher effects for four different teacher types: math, English, science and social studies. The primary contribution of the paper is that it allows for a full treatment of the joint-production environment in secondary school. Although the teacher-quality literature has generally assumed that same-subject teacher quality influences student performance (i.e., math teachers affect math performance and English teachers affect reading performance), it has also implicitly assumed that off-subject teacher quality does not (i.e., math teachers do not affect reading performance). This latter assumption lacks empirical support and by relaxing it, I find strong evidence of joint production in secondary school. In each

tested subject, distributional shifts in teacher quality for both same-subject and off-subject teachers can have large effects on student performance.[3]

This analysis also provides a unique opportunity to compare the effects of distributional shifts in teacher quality at elementary and secondary schools.[4] This comparison offers insight into the teacher-quality allocation problem across schooling levels. By a sizeable margin, the influence of variation in teacher quality on student performance is larger at the elementary level.

I.II. The Educational Production Function

I evaluate teacher quality based on secondary-school students' test-score outcomes in math and reading. Student achievement in any given year is the result of a cumulative set of inputs from families, peers, communities and schools. Because data on the complete histories of students are unavailable, researchers have focused on estimating educational production in terms of value added. The general value-added framework explains current performance as a function of current inputs while controlling for past performance:[5]

$$Y_{isjt} = f(Y_{isj(t-1)}, a_i, X_{it}, \delta_s, S_{it}, C_{it}, \theta_{1j}, \theta_{2j}, ..., \theta_{Kj})$$

[3] Performance in math and reading depends on multiple teacher inputs; however, all four teacher types do not belong in each specification. See Section V for details.
[4] Through comparison with Koedel and Betts (2007), which focuses on elementary-level educational production.
[5] A specific form of the general value-added model is the gainscore model in which researchers subtract past performance from current performance and treat the gain in performance as the outcome of interest.

Here, Y_{isjt} represents the educational outcome of interest for student *i* at school *s* with teacher-set *j* in year *t*, α_i represents observed and unobserved time-invariant student characteristics, X_{it} is a vector of time-varying observable student characteristics, δ_s represents observed and unobserved time-invariant school characteristics, S_{it} is a vector of observed time-varying school characteristics, C_{it} is a vector of time-varying observable classroom characteristics, and θ_{kj} measures the quality of teacher k (who is part of teacher-set j). The above formulation incorporates a vector of teacher effects representing the inputs of multiple teachers per year and the subscript j corresponds to a set of teacher effects.[6]

I empirically examine three important questions relating teacher quality to student performance based on the production function above. First, to what extent is secondary-school educational output the result of multiple teacher inputs? Second, which teacher inputs are most important in the production of which outputs? Third, in cases where educational output is produced by multiple teacher inputs, how does teacher quality across subjects interact in the production function? For this latter question, I initially assume that teacher quality across subjects does not interact at all. Later, I relax this assumption and evaluate the potential for teacher interactions to influence student performance.

[6] In the elementary case, this set would consist of just one teacher.

Consider four teacher types: math, English, science and social studies.[7] Index math teachers from j = 1,…,J; English teachers from p = 1,…,P; science teachers from q = 1,…,Q; and social studies teachers from r = 1,…,R. For student *i* who has the *j*th math teacher, the *p*th English teacher, the *q*th science teacher and the *r*th social studies teacher; the set of teacher effects influencing her performance can be defined as $(\theta_j, \theta_p, \theta_q, \theta_r)$ where θ_j indicates the quality of math teacher *j*, θ_p indicates the quality of English teacher *p*, and so on. I focus on the effects of these four teacher types in determining student test-score performance on math and reading standardized tests.

I estimate teacher fixed effects using a within-school and student value-added specification in the reduced form:

$$\textbf{(1)}\quad TestScore_{ist}^{jpqr} = \alpha_i + TestScore_{is(t-1)}^{jpqr}\psi + X_{it}\gamma + D_{it}^{school}\delta_S + S_{it}\rho + C_{it}\eta + D_{it}^{J(math)}\theta_J + D_{it}^{P(eng)}\theta_P + D_{it}^{Q(sci)}\theta_Q + D_{it}^{R(soc)}\theta_R + \varepsilon_{it}$$

Teachers are indexed by subject as indicated above and denoted by superscripts. All of the explanatory variables are defined above and a detailed list of the sets of controls in each vector is in Table 1.1. Vectors of indicator variables for schools and teachers are denoted by a "D" and are appropriately labeled. This

[7] These four teacher types are the most common in San Diego high schools and arguably most relevant for evaluating cognitive performance. Among the remaining teacher types that are omitted from this analysis, some of the more common teachers include language teachers and art teachers. The class-taking behavior of my student sample is detailed in Section IV.

specification allows for joint production among teachers in high school by allowing for multiple teachers to affect student outcomes in both math and reading.[8]

On the one hand, the within-school and student specification in equation (1), which includes school- and student-level indicator variables, minimizes omitted variables bias due to unobserved heterogeneity in school quality and student ability. However, on the other, it ignores any between-school variation in teacher quality. To the extent that teachers vary in quality across schools, the within-school and student estimates will understate the total variance of high school teacher quality. In Appendix 1.D, I evaluate the sensitivity of my estimates to alternative specifications that allow for between-school variation in teacher quality.[9]

To control for the variety of different types of classes that students take in high school, the vector of classroom controls (C_{it}) includes indicator variables for the subjects and levels of subjects that each student takes in each year (e.g., algebra or geometry, regular or honors English, etc.). This prevents variation in subject material from being attributed to variation in teacher quality and means that teacher quality is measured within subject and subject level. To address the issue of peer effects, the model includes controls for the year (t-1) achievement of classroom-level peers for

[8] However, the model in (1) does not allow for complementarities between teachers as would be the case if the various teacher indicator variables were interacted. I evaluate the importance of teacher interactions below.

[9] Appendix 1.D shows that between-school variation in teacher quality among San Diego high schools may be non-negligible. Therefore, I expect my analysis to understate variation in teacher quality in secondary school to some degree.

each student's math and English classrooms.[10] Also, note that the effects 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. Finally, I control for class size to prevent variation in class size from being misinterpreted as variation in teacher quality.[11]

I adopt the method of Anderson and Hsiao (1981) to estimate the model in (1). This 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 $(TestScore^{jpqr}_{is(t-1)} - TestScore^{jpqr}_{is(t-2)})$ with $(TestScore^{jpqr}_{is(t-2)})$. 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), I use the first-differenced error terms within students to test for serial correlation between the ε_{it}'s and find that this primary assumption is upheld.[12] The first-differenced version of equation (1) is detailed below:

---

[10] I also run models that include peer and class-size effects for social studies and science classrooms, although these models are complicated by the fact that not all students take science and social studies classes in each year. Regardless, the inclusion of these additional controls has a negligible effect on results.

[11] Controls are included for math and English class sizes only. Class-size controls have a negligible effect on teacher quality estimates.

[12] 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-

$$
\begin{aligned}
(TestScore_{ist}^{jpqr} - TestScore_{is(t-1)}^{jpqr}) = \quad & (\alpha_i - \alpha_i) + (TestScore_{is(t-1)}^{jpqr} - \widehat{TestScore}_{is(t-2)}^{jpqr})\psi \\
& + (X_{it} - X_{i(t-1)})\gamma + (D_{it}^{school} - D_{i(t-1)}^{school})\delta_S + (S_{it} - S_{i(t-1)})\rho + (C_{it} - C_{i(t-1)})\eta \\
& + (D_{it}^{J(math)} - D_{i(t-1)}^{J(math)})\theta_J + (D_{it}^{P(eng)} - D_{i(t-1)}^{P(eng)})\theta_P + (D_{it}^{Q(sci)} - D_{i(t-1)}^{Q(sci)})\theta_Q \\
& + (D_{it}^{R(soc)} - D_{i(t-1)}^{R(soc)})\theta_R + (\varepsilon_{it} - \varepsilon_{i(t-1)})
\end{aligned}
$$

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.[13] This first-differenced model will produce unbiased coefficient estimates while at the same time accounting for a wide set of covariates.[14]

I.III. Methods

Because my analysis includes over 1000 high school teachers, tables displaying individual coefficient estimates for teachers would be difficult to interpret. Instead, I describe the variance of the teacher-quality distribution. First, I perform Wald tests for the joint significance of the sets of teacher fixed effects using different versions of equation (1). These tests evaluate the statistical significance of variation in teacher quality as a determinant of educational output and are of the form:[15]

---

differenced equation is estimated, I estimate that the serial correlation between the first-differenced error terms to be -0.45.

[13] The period (t-2) test-score level is a powerful instrument: t-statistics on the period (t-2) test-score are greater than 50 for each of the first-stage models.

[14] Robust standard errors for all 2SLS coefficients in this model were used. In addition, the differenced error terms are serially correlated for students with more than one first-differenced equation in the model (that is, at least 4 test-score records) per the previous footnote. I structurally enforced this property of the error term in the variance-covariance matrix for relevant students.

[15] The examples in this section are for math-teacher effects per the previously defined notation. Statistical analysis for all other teacher types is similarly performed.

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_J = \bar{\theta}$$

(2) $$W = (\hat{\theta} - \bar{\theta}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \bar{\theta}\ell_J)$$

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.[16] Under the null hypothesis, W is distributed $\chi^2_{(J-1)}$.

Although the Wald test allows for the identification of statistical significance, it does not provide an estimate of the magnitude of the variance of teacher quality. To determine *economic significance*, I empirically estimate the variance of teacher quality. First, I calculate the total fixed-effects variance for each teacher type from the models of student achievement for math and reading. For math teachers, this variance is:

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

Each fixed-effect coefficient is comprised of two components - one consisting of the true signal of teacher quality and the other of estimation error, $\hat{\theta}_j = \theta_j + \lambda_j$.

[16] The variance matrix used in my 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 my results.

Equation (3) overstates the variance of teacher quality because it includes the variance of the estimation error. I 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, I first assume that $Cov(\theta, \lambda) = 0$.[17] This allows for the total variance of teacher fixed effects to be decomposed as follows:

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

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

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

Equation (5) weights the total fixed-effects variance by the estimation error variance on a coefficient-by-coefficient basis. See Appendix 1.B for more detail.

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

[17] 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.

estimated to be *A* then the magnitude of the variance of the teacher-quality signal is estimated by:

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

I use the estimates of $Var(\theta)$ from the quality distributions of the different teacher types to evaluate the effects of distributional shifts in teacher quality on student performance.

I.IV. Data

This study uses panel data from the San Diego Unified School District (SDUSD) following high school students and teachers over time.[18] SDUSD is the second largest school district in California (enrolling approximately 141,000 students in 1999-2000) and the student population is approximately 27 percent white, 37 percent Hispanic, 18 percent Asian/Pacific Islander and 16 percent black. 28 percent of the students at SDUSD are English Learners, and 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 at SDUSD trailed very slightly behind the national average in reading in 1999-2000. On the contrary, SDUSD students narrowly exceeded national norms in math.[19]

[18] The data used for the dropout analysis will be discussed separately below.
[19] District characteristics summarized from Betts, Zau and Rice (2003).

The test-score data are from the Stanford 9 test spanning the school years from 1997-98 through 2001-02. Students are tested from the eighth through the eleventh grade.[20] Students and teachers are linked at the classroom level and an extensive set of school, student and classroom characteristics is available. Table 1.1 details the data used in this analysis.

There are 16 standard high schools at SDUSD and a handful of other schools that offer secondary-level instruction (either charter schools or schools of an atypical grade structure - for example, grades 7 – 12 or K – 9). Among these 16 standard high schools, enrollment in 1999-2000 ranged from 849 to 2,945 students. Among the charter and atypical schools, secondary-level enrollment ranged from 26 to 1,039 students. The data for this study are based primarily on students from the standard high schools at SDUSD. However, some students from atypical or charter schools are also included.[21]

The modeling structure in equation (1) requires that all students used in this analysis have at least three contiguous test-score records at SDUSD (which covers a geographically large area). Students who do not satisfy this criterion are omitted from this study. Appendix 1.A provides summary statistics showing that the final student sample is slightly advantaged relative to the entire student population at

[20] Eighth-grade test-scores are used only as (t-2) explanatory variables in the final models.
[21] Data were not available for all charter and atypical schools. The model includes school fixed effects to control for heterogeneity in school types.

SDUSD but is generally representative. In subsequent analysis, I will show that the omission of student fixed effects from the student-achievement specification results in inaccurate estimates of teacher fixed effects, justifying the approach.

I also require that each student have both a math and English teacher in each year in which his or her data are used. This facilitates a straightforward comparison between math and English teachers - it ensures that they are evaluated using the same student set. The benefit of being able to directly compare the relative importance of these separate educational inputs seems to outweigh the cost of lost data due to this restriction because the sample size remaining after imposing this restriction is still quite large.[22] Table 1.2 details the class-taking behavior of my student sample by grade level.

Despite the restrictions imposed on the student sample, the data still include over 53,000 test-score records from over 15,000 different students. Because my student sample is likely to be more homogeneous than the general student population, my results may understate the variance of teacher quality.[23]

---

[22] I exclude 3.8 percent of the student sample because they are not assigned to a math class in at least one year and 8.7 percent of the student sample because they are not assigned to an English class in at least one year. The latter group is peculiar because the general high school curriculum is such that each student should take English each year, including English learners. Some of these omissions may reflect students moving in and out of the district over time. Others may be due to missing data.

[23] Appendix Table 1.A.1 shows that the student sample used here is generally representative of the student population at SDUSD. The primary difference is that the student population used here outperforms the entire population in terms of test scores.

I also impose participation restrictions on teachers. By analogy to Kane and Staiger's analysis of school quality (2002), I expect sampling variation to have a significant impact on estimated teacher effects. Thus, I require that teachers have at least 20 student-years of data to be included in the analysis.[24] After imposing this restriction, just over 1000 high school teachers are included.[25] Appendix 1.A provides summary statistics for the teacher sample used in this analysis.

I.V. Results – Statistical Significance of Teacher Effects

I perform Wald tests for the statistical significance of variation in teacher quality as a determinant of student outcomes.[26] The results from these tests indicate which teacher inputs affect which test-score outputs in secondary school. Tables 1.3 and 1.4 present results from these tests for the math and reading models, respectively, based on the student-achievement specification in equation (1). In both cases, I begin with basic models that include only same-subject teachers and subsequently consider the inclusion of all possible teacher combinations.

Tables 1.3 and 1.4 show that variation in teacher quality among same-subject teachers is a statistically significant determinant of students' test-score outcomes in

[24] That is, 20 student-years of data from the restricted pool of students. The results presented in this paper are not sensitive to a reasonable range of adjustments to this threshold.
[25] I estimate effects for 346 English teachers, 269 math teachers, 202 science teachers and 184 social studies teachers.
[26] The magnitudes of the variances of math and reading test scores are very similar. The raw standard deviations of the math and reading test-score distributions are 35.7 and 37.2, respectively. The standard deviations of the residuals after taking out the within school and student variation are 15.0 and 13.6 for math and reading.

both math and reading for all relevant specifications.[27] In the reading models, variation in math-teacher quality is also always a significant determinant of student outcomes. However, the same is not true for variation in English-teacher quality in the math models. Finally, whereas variation in teacher quality among social studies teachers seems to generally affect student outcomes in both math and reading, variation in teacher quality among science teachers does not affect performance in either subject.

Recall that the teacher effects enter into equation (1) linearly. However, teacher quality across subjects may interact in the production function. Based on the results from the Wald tests above, I test to see if teacher-interaction effects belong in the math and reading models. For math production, I add interaction terms between math and social studies teachers to model (4) in Table 1.3. For reading, I add interactions between English and math teachers, English and social studies teachers, and math and social studies teachers to model (5) in Table 1.4.[28]

For math output, the Wald test for the joint significance of the set of interaction terms between math and social studies teachers indicates that these interactions do not belong in the model (the p-value from this test is 0.70). Similarly, for reading output, interactions between English and social studies teachers and math

[27] The exception to this is in the math models that include English-teacher indicator variables. In each of these models the set of math-teacher coefficients is jointly insignificant. However, the set of English-teacher indicator variables clearly does not belong in the math-achievement model.
[28] To maintain consistency in my data inclusion restrictions, I require teacher interactions to affect at least 20 students to be included into the model.

and social studies teachers are also jointly insignificant (p-values of 0.95 and 0.97 respectively). However, interactions between English and math teachers in the reading model are significant at the 1 percent level of confidence. Furthermore, the inclusion of the math and English teacher interactions into the reading model lowers the Wald statistic for the joint significance of the social studies teacher indicator variables to the point where they are no longer identified as statistically significant determinants of reading performance.[29] Therefore, the final reading-achievement specification is *not* model (5) in Table 1.4, but instead includes indicators for just math and English teachers as well as interaction terms between these two teacher types. It excludes both science and social studies teachers. Table 1.5 details this final reading-achievement specification.

I.VI. Results –The Magnitude of the Variance of Teacher Quality

VI.A Math Analysis

In this section teacher quality is measured strictly in terms of student math performance regardless of each teacher's primary subject of instruction. I start with the "basic" math model that ignores the possibility of joint production among teachers and includes only math-teacher effects (model (1) in Table 1.3). To the extent that

[29] The p-value on this new Wald statistic for the inclusion of the social studies teacher indicator variables is approximately 0.90. This result is maintained even if all interactions involving social studies teachers are removed from the model (that is, it is the inclusion of the English-math teacher interactions that causes the Wald statistic to fall to the point of statistical insignificance). It may be that, given a five- or six-class schedule, students' social studies teachers are strong predictors of their math and English teacher combinations. Because of this, I also test for the statistical significance of math and English teacher interactions in the math model despite the results from the Wald tests in Table 1.3. These interaction effects are jointly insignificant in the math specification and the social studies teacher indicator variables retain their statistical significance.

students are non-randomly assigned to social studies teachers, this basic model will produce biased estimates of teacher fixed effects. Therefore, I also estimate the full model of student math achievement that includes fixed effects for social studies teachers (model (4) in Table 1.3).

For each teacher type and in each model, I report the unadjusted raw variance of teacher fixed effects and also the adjusted variance of teacher quality as estimated by equation (6). For ease of interpretation, results are presented as the ratio of the standard deviation of the teacher quality distribution to the weighed average of the within-grade standard deviations of test scores (where the weights correspond to the sample size in each grade).[30] For example, in the basic model, Table 1.6 indicates that a one-standard-deviation increase in math-teacher quality (adjusted) corresponds to a 0.08 average within-grade standard deviation improvement in student test scores.

The results from the full math achievement model in Table 1.6 indicate the tradeoffs in teacher quality across subjects required to maintain a given level of achievement growth. Because the achievement-growth isoquants of the math educational production function are roughly linear in teacher-quality space (per the interaction-effect Wald tests in the previous section), it is straightforward to calculate their slope (the marginal rate of technical substitution). For example, an equivalent

[30] This metric is chosen because it allows for the most straightforward comparison of results across studies. However, it may be slightly misleading because the model specification in equation (1) does not allow across-school or across-student variation in teacher quality while this metric measures teacher quality relative to the *total* variation in test-scores. Nonetheless, the estimates are sizeable.

gain in math test-scores can be achieved by either a one-standard-deviation increase in math-teacher quality or a 1.05-standard-deviation increase in social studies teacher quality.

Table 1.6 indicates that the tradeoff in teacher quality between math and social studies teachers required to maintain a given level of achievement growth, *measured by standard deviations of each teacher-type's respective quality distribution*, is roughly one-to-one. However, this does not mean that teacher quality across subjects measured in levels, which I cannot observe, will trade off at the same rate. For example, if we assume that math-teacher quality is more important in determining math outcomes than is social studies teacher quality, Figure 1.1 may simply reflect the fact that there is more heterogeneity in quality among social studies teachers. In this case, a one-standard-deviation improvement in teacher quality among social studies teachers would represent a larger absolute change. There is suggestive evidence from the credentialing process indicating that math teachers may indeed be a more homogenous group than social studies teachers. For example, the first time pass rate for the math-credentialing exam in California is just 29.2 percent. For the social studies exam, the pass rate is over 62 percent.[31, 32]

[31] Passing rates from *Report on Passing Rates of Commission-Approved Exams for 2000-01 to 2004-05* from the California Commission on Teacher Credentialing released in April 2006 and are for California as a whole. Reported passing rates are from July 2003 through July 2005 and therefore are not directly applicable to the teacher set used here. However, other sources confirm a similar relationship between passing rates on the different exams in the 1990s.

[32] Another factor that may explain the results in Table 1.6 is differences in the rigidity of curriculums across math and social studies teachers. For example, a high-school economics teacher can teach a mathematical economics class or a non-mathematical economics class, whereas a math teacher has less

I consider the extent to which the variation in math-teacher quality estimated in Table 1.6 is linked to observable teacher qualifications by running another regression in which I omit all of the teacher indicator variables and instead include controls for math teachers' experience, credentialing, education levels and whether or not each math teacher has an undergraduate degree in mathematics.[33] None of these observable teacher qualifications have statistically significant effects in the model.[34] Additionally, the effects implied by the point-estimates on these variables are very small.

The aggregated effect of a one-standard-deviation improvement in teacher quality across both relevant teacher types in math is equivalent to a 0.13 within-grade standard deviation improvement in test scores.[35] Compared to estimates of the effects of other educational inputs on secondary-school output throughout the literature, this implies that teacher quality is likely to be the most effective policy-relevant tool at the

---

discretion in curriculum. Variation in curriculums across social studies classes will be captured by the teacher effects, perhaps rightfully so.

[33] For experience, I estimate models that allow experience to enter linearly (up to 10 years of experience) and also models that include indicator variables for teachers with two or less years of experience. I also control for whether teachers have a master's degree and whether they are fully credentialed.

[34] The highest p-value for any of these coefficients is 0.26 and is for the coefficient on the new-teacher indicator variable.

[35] The estimates here are somewhat smaller than estimates reported by Aaronson, Barrow and Sander (2007). This may have to do with differences in the testing instruments employed to estimate teacher effects in the two studies. Aaronson, Barrow and Sander report that in their study, student test-score growth differs substantially by students' initial achievement levels and that high-achieving students experience much larger test-score gains from 8th to 9th grade (the grades studied by these authors). In the presence of positive student-teacher matching, this would be expected to inflate the variance of their estimated teacher effects. Nonetheless, my estimates confirm the general result from Aaronson, Barrow and Sander (2007) that variation in teacher quality is an important determinant of student outcomes in secondary school.

disposal of administrators.[36] For example, one of the more popular policy interventions discussed within the educational community is class-size reduction. Results from independent studies by Betts, Zau and Rice (2003) and Rivkin, Hanushek, and Kain (2005) indicate that variation in class size has no effect on student achievement as students move beyond elementary school.[37]

VI.B Reading Analysis

In this section teacher quality is measured exclusively in terms of student reading performance. I begin with the basic reading model in which joint production among teachers is ignored and student performance is attributed solely to variation in English-teacher quality. Next, I estimate the complete reading achievement model as described in model (9) in Table 1.5.

Similarly to the math analysis, the results from the full reading achievement model in Table 1.7 indicate the tradeoffs in teacher quality across subjects required to maintain a given level of achievement growth. However, unlike for math, the reading production function is not strictly linear in teacher-quality inputs.

The nonlinearity between math- and English-teacher quality may represent some combination of the effects of teacher cooperation and teacher matching, or the

[36] The body of literature that estimates the effects of observable educational inputs on student outputs is vast. See Hanushek (1986, 1996) for literature surveys.
[37] Estimates from my analysis confirm these authors' findings.

effect of the compounding of quality across subjects (i.e., increasing or diminishing returns). Because the data do not contain direct information on teacher quality, these effects are difficult to disentangle.[38] However, by examining the interaction effects for teachers of different estimated quality levels, it is possible to at least partially identify the extent to which the teacher interactions reflect increasing or decreasing returns to teacher quality inputs.

To evaluate the returns to scale of teacher-quality inputs across subjects I start by dividing the English- and math-teacher coefficients from the full reading model into two separate vectors. Within each vector, I rank teachers from 1 to P and 1 to J, respectively, based on their value-added coefficients and assign them to quintiles based on their rankings, where quintile-5 teachers are those with the highest value added. Table 1.8 shows the average interaction effect experienced by students whose teachers are from any given quintile set, where a quintile set is defined by the pair of quintile rankings for each student's English and math teachers (i.e., the set (1,4) would indicate an English-teacher quintile ranking of "1" and a math-teacher quintile ranking of "4"). The results in Table 1.8 are presented in terms of the same weighted average of within-grade test-score standard deviations as the results in Table 1.7.[39]

[38] Teacher quality is estimated from the student-achievement specifications based entirely on student outcomes.

[39] Because Table 1.8 displays average effects, the estimates are not "adjustable" as are the variance estimates in Table 1.7. However, if the estimation error for the teacher-interaction coefficients is independent of teachers' quintile rankings, the estimation error in the reported interaction averages in each cell of Table 1.8 should be zero, on average.

The estimates in Table 1.8 show that one source of the nonlinearity in the reading model is diminishing returns to teacher quality inputs across subjects. For example, the table shows that students who are taught by high-quality teacher sets, on average, do not experience the full performance gain that would be implied by the simple sum of their teachers' effects.

Because the production of reading output is characterized by diminishing returns to teacher quality inputs (and possibly teacher cooperation and/or teacher matching as well), estimating the effect of an improvement in teacher quality on student performance is less straightforward than in the math analysis. However, generally speaking, the estimates in Tables 1.7 and 1.8 indicate that the effect of a one-standard-deviation improvement in math- or English-teacher quality can have a substantial effect on student performance. For example, for a student being taught by a 3rd-quintile teacher in both English and math, the effect on performance of a one-standard-deviation improvement in teacher quality in either subject (to the 5th quintile) would be equivalent to the full effect detailed in Table 1.7.

Finally, I evaluate the extent to which variation in English-teacher quality can be explained by observable teacher qualifications. In the basic model, I omit all English-teacher indicator variables and instead include controls for English teachers' experience, credentialing, education levels and whether or not each teacher has an undergraduate degree in English. Only the coefficient on the indicator variable for

whether an English teacher has a master's degree is statistically significant and the implied effect is quite small.[40] As in the math analysis, compared to the larger educational production literature that considers the effects of observable inputs such as spending per pupil and class-size reductions on student performance, the reading analysis indicates very large teacher-quality effects that are virtually unrelated to observable teacher qualifications.

I.VII. The Superstar Teacher Hypothesis

Sections V and VI show that math teachers affect both math and reading achievement. Does this imply that some math teachers are so great that they positively affect both math and reading performance, the proverbial "superstar teacher" effect, or so bad that they negatively affect performance in both subjects? Or does this instead imply that math teachers are making tradeoffs that influence their effectiveness in math and reading and that, generally speaking, performance in one subject is obtained at a cost in the other? This question can be addressed by analyzing the correlation of math-teacher effects across subjects. A strong positive correlation would confirm the superstar teacher effect.

Define $\underline{\hat{\theta}}_m$ as the vector of estimated math-teacher coefficients from the full math model and $\underline{\hat{\theta}}_r$ as the vector of estimated math-teacher coefficients from the full

[40] Having an English teacher with a master's degree is estimated to improve performance by .01 within-grade standard deviations of the test.

reading model. The correlation between these two vectors is 0.31. 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), by assuming that the correlation of true teacher quality across subjects is the same for all teachers, I can get an idea of the direction of the bias introduced by the measurement error. Measurement error will be smaller for teachers with a greater number of student-year observations. Therefore, I 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 bias from $\underline{\lambda}_m$ and $\underline{\lambda}_r$ on the initial quality-correlation estimate. The estimated correlation coefficient from the selected subset of teachers is higher than its counterpart from the full teacher set. Thus, measurement error is biasing the estimate of the correlation of teacher quality across subjects toward zero and the initial estimate of the correlation between $\underline{\hat{\theta}}_m$ and $\underline{\hat{\theta}}_r$, 0.31, can be treated as a lower-bound estimate of the correlation of math-teacher quality across subjects.

To estimate an upper bound on the correlation of math-teacher quality across subjects, I estimate the correlation between $\underline{\theta}_m$ and $\underline{\theta}_r$ under the assumption that the true correlation between $\underline{\lambda}_m$ and $\underline{\lambda}_r$ is zero (See Appendix 1.C for details). This upper-bound estimate does not exclude the possibility that the correlation of math-

teacher quality across subjects is equal to 1. The bounded estimate of the correlation of math-teacher quality across subjects (0.31 to 1.00) provides support for the superstar teacher hypothesis.[41]

I.VIII. Specification Checks

I now examine the robustness of the teacher fixed effect estimates to various alternative models. Table 1.9 shows four separate value-added specifications for the model of student achievement. The first column shows the full model estimated in equation (1) and columns 2 through 4 show restricted models. 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.[42]

The models used in Table 1.9 are of the "basic" form (corresponding to the first vertical panels of Tables 1.6 and 1.7), meaning that only same-subject teachers

[41] The identification of the mechanism by which math teachers affect reading performance is beyond the scope of this project. It may be that math teachers directly influence reading skills through their teaching (e.g., by focusing on word problems that improve reading comprehension). Alternatively, it may be that math teachers are particularly important to student confidence and motivation. In the education literature, there is a term for the distress to students caused by math – "Mathematics Anxiety" (see, for example, Hembree, 1990). Additionally, popular media has argued that algebra is a particularly devastating subject for some students' confidence levels (Helfand, 2006).

[42] P-values from Wald tests of the null hypothesis that the coefficients on the omitted variables in the restricted models are zero are less than 0.01 for all student-level covariates and the set of school-level covariates and school fixed effects in each specification (variable groups B and C in Table 1.9). I 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 my 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 my major findings are generally robust to models of student achievement that are not first-differenced (see Appendix 1.D for more details). 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 1.9) and merits additional attention in future research.

are included into the specification for both math and reading.[43] For each restricted specification, I estimate the vector of teacher fixed effects for the relevant teacher type (math or English) and compare it to its analog from the complete specification in column 1 by reporting the correlation between the vectors. This is one measure of the magnitude of the effect of the omitted variables bias in the different models insofar as this bias affects the estimated teacher coefficients. The lower the correlation between the two vectors of teacher fixed effects, the larger the omitted variables bias.

Columns 2 through 4 of Table 1.9 show that it is important to include each major variable group in the model of student achievement in order to obtain unbiased estimates of teacher fixed effects. However, estimating the variance of teacher quality in the absence of some of the variable groups in Table 1.9 may be of interest in some cases. For example, the set of school- and classroom-level variables may be strong predictors of student performance, but one source of their predictive power may be teacher sorting. To the extent that this is the case, we may be interested in teacher-quality variance estimates from models that exclude these variables in order to capture between-school as well as within-school variation in teacher quality. In Appendix 1.D, I consider the sensitivity of the previously reported estimated variances of teacher quality to the alternative specifications detailed in Table 1.9, and also to an equivalent set of specifications based on test-score levels. This analysis indicates that between-school variation in teacher quality across San Diego high

[43] An analogous exercise using the full math and reading models with all relevant teachers offers qualitatively similar results.

schools may be non-negligible and because of this, the estimates presented in this study potentially understate the total variance of secondary-school teacher quality.

I.IX. Test Scores and Teacher Accountability in Secondary School

The degree to which value-added modeling can be used as a tool for determining teacher accountability in secondary school may be the most timely policy-related question in this analysis. I address two accountability-based issues here. First, how much information about outcome-based teacher quality can be obtained from noisy teacher-fixed-effect estimates? Second, and specifically relevant for the high school setting, how do decisions regarding which teacher-types (i.e., math, English, science, social studies) to include in the model of student achievement affect teacher rankings based on value added?

To evaluate the precision of the individual teacher coefficients, I consider the portion of these coefficients that, on average, represents the true signal of teacher quality. The greater the signal portion of these coefficients, the more useful they will be for teacher evaluation. Section III proposes that the signal and noise components of the variance for any *set* of estimated teacher coefficients can be separated. If the individual teacher coefficients, on average, are representative of the entire group, I can evaluate the average signal-to-noise ratio that characterizes each of these individual coefficients.

In the full math achievement model described in Table 1.6, the variance decomposition in equation (6) indicates that 23 percent of the total fixed-effects variance for math teachers represents the signal of true teacher quality suggesting that, on average, the individual math-teacher coefficients are themselves 23 percent signal. Similar estimates imply that the individual social studies teacher coefficients from the same model are 35 percent signal. In the complete reading model from Table 1.7, the English and math-teacher coefficients are, on average, 32 and 35 percent signal respectively. On the one hand, the average signal portions of the individual teacher coefficients seem low, suggesting that the incorporation of value-added modeling into teacher evaluations should be cautiously approached. However, on the other, value-added estimates may still represent a marked improvement over the measures most commonly used to determine teacher recruitment, retention and salaries (e.g., teachers' education levels, credentials and experience).[44]

If policymakers were to incorporate value-added estimates into the evaluations of secondary-school teachers, the strategy for modeling student achievement would be a second relevant concern. For example, in the calculation of value added for math teachers, should the model of student math achievement include the effects of only math teachers or should it include the effects of social studies

[44] For example, my analysis in Section VI shows that observable teacher qualifications are almost entirely unrelated to student performance in both math and reading. For additional evidence see Aaronson, Barrow and Sander (2007), Angrist and Guryan (2003), Betts (1995), Betts, Zau and Rice (2003), Hanushek (1986, 1996), Kane, Rockoff and Staiger (2006), Koedel and Betts (2007).

teachers as well (per model (4) in Table 1.3)?[45] Because such a decision may involve political as well as economic considerations, the more important question is perhaps whether such a choice will make a significant difference in determining teacher rankings.

I consider the hypothetical example of an accountability system in which math teachers are evaluated based on their ranking in terms of math value added and English teachers are evaluated based on their ranking in terms of reading value added.[46] First, for math teachers, I estimate the basic model that assumes only math teachers affect student math performance (Table 1.6, panel 1). I keep the vector of math-teacher coefficients and rank them from 1 to J, 1 being the lowest and J being the highest. Next, I estimate the full model of student math achievement that also allows for social studies teachers to also affect math performance (Table 1.6, panel 2). From this model, I keep just the vector of coefficients for math teachers and again rank them from 1 to J.

For each vector of math-teacher coefficients, I divide teachers into quintiles based on their value-added rankings, where quintile-5 teachers are those with the

[45] A related question is: How should we determine which teachers will be held accountable for gains in student achievement in which subjects? The empirical results presented above indicate that multiple teachers do appear to play roles in determining both math and reading achievement for students. However, the benefits of cross-subject accountability would have to be weighed against the costs which may come in the form of subject material tradeoffs. For example, if social studies teachers were to be rewarded for student math performance, would they displace important social studies material to improve math test scores? An additional concern is the introduction of free-riding opportunities.

[46] I will assign each teacher an overall quality measure despite the fact that performance is measured within schools. If there is significant between-school sorting in terms of teacher quality, the rankings I assign will be less comparable across schools.

highest value added. Table 1.10 compares the "stability" of these quintile assignments across the different models of student achievement. Each cell entry in Table 1.10 indicates the percentage of teachers who fall into a given quintile set, where a quintile set is defined by the pair of quintile-rankings for a given teacher in both models. The vertical dimension represents teachers' quintile rankings from the basic model and the horizontal dimension teachers' rankings from the full model. The correlation between the two vectors of math-teacher coefficients is 0.95.

If math teachers' value-added coefficients were independent of social studies teachers' value added and if the inclusion of the social studies teacher indicator variables into the model did not introduce any additional noise, the diagonal entries of Table 1.10 would all equal 100 percent and the off-diagonal entries would all equal zero. Although this is certainly not the case in the center of the matrix, the corners of the matrix indicate that the best and worst math teachers are generally identified regardless of whether the social studies teacher indicator variables are included or not. Importantly, it is precisely these teachers who we would expect to target in an accountability system. Thus, for relevant teachers, Table 1.10 implies a relatively low omitted variables bias generated by the omission of social studies teachers in the basic model of student math achievement and indicates that a simple teacher-accountability system that rewarded math teachers based on such a model should perform relatively well. Put differently, Table 1.10 shows that objections to the assignment of teacher accountability in high school based on the contamination of

teacher effects across subjects, at least among the highest- and lowest-ranked teachers, would be largely misguided.

Next, I perform an analogous exercise for English teachers in the reading achievement specification. In this case, I compare the basic model that assumes that only English teachers affect student reading performance to the full model detailed in Table 1.5 (including English and math teachers as well as interactions between the two). The quintile stability results are displayed in Table 1.11. For this analysis, the correlation between the two vectors of English-teacher coefficients is 0.87.

The results in Table 1.11 are quite similar to those in Table 1.10. For English teachers, switching between the models of student achievement has a slightly larger effect on teachers' rankings. However, the best and worst teachers are still consistently identified. As in Table 1.10, teachers ranked in the top and bottom quintiles in the basic model are particularly likely to be ranked in the same quintile in the full model. Furthermore, for top- and bottom-quintile teachers as identified by the basic model, 93 and 98 percent, respectively, are identified as being in the top or bottom two quintiles in the full model.

The evidence here supports previous work indicating that value-added modeling is most consistent in identifying the best and worst teachers regardless of the type of distortion introduced for comparison (e.g., adjustments in time, student

sample, or in this case, model completeness).[47] Value-added modeling will be most useful in an accountability system that focuses on these types of teachers, which is what seems most reasonable.[48] Finally, note that value-added estimates may be better suited as part of a larger and more comprehensive system of teacher accountability as opposed to a stand-alone measure of quality. Specifically, incorporating value-added estimates with other measures of teacher quality that would be expected to have uncorrelated errors will improve the quality signal. For example, evaluations based on classroom observation and/or principals' recommendations (not test-score based) may represent prime candidates to combine with value-added modeling to reduce erroneous rewards or sanctions.

I.X. The Effects of Variation in Teacher Quality Across Schooling Levels

A question of great practical importance to the educational production literature is whether teacher quality is more valuable as a resource at the elementary or secondary level. Unfortunately, I cannot directly observe the effects of absolute teacher quality. However, I can observe student-performance responses to variation in teacher quality and answer a related question: Does variation in teacher quality have larger effects in elementary or secondary school?

---

[47] Also see Koedel and Betts (2007) and Aaronson, Barrow and Sander (2007)

[48] For example, a system that rewards teachers ranked around the 60th percentile seems much less practical than one that rewards teachers ranked around the 80th or 90th percentile in terms of performance.

A note of caution in this analysis is warranted. We lack important information about the heterogeneity in teacher quality across schooling levels. There is suggestive evidence that the elementary-level teacher population is a more heterogeneous group than the secondary-level population, particularly when looking within-subject in secondary schools. For example, the pass rates for credentialing tests are much higher for elementary-school teachers than they are for subject-specific secondary-school teachers (particularly math teachers). Also, elementary teachers are much more likely to have general undergraduate degrees (e.g., education, social science). Although the subjects of teachers' undergraduate degrees are, at best, weakly linked to student performance, the presence of a large population with general degrees may imply more heterogeneity in the group as a whole. Because the reported variance estimates incorporate heterogeneity, we should perhaps expect more variation in teacher quality among elementary teachers from the outset regardless of any differential student responses to differences in teacher quality across schooling levels.

Table 1.12 compares estimated variances of teacher quality for elementary- and secondary-school teachers in math and reading. The elementary-level results are from Koedel and Betts (2007) and are particularly relevant for comparison here because they are generated using the same standardized test (the Stanford 9), the same timeframe, the same school district and the same general value-added specification as

the results from this analysis.[49] Note that elementary-school teachers have the same students for the entire day whereas secondary-school teachers have each student for just one hour. Therefore, for secondary school, I present two different sets of estimates. The first column of secondary-school estimates shows the effect of a one-standard-deviation change in teacher quality for same-subject teachers. The second column of secondary-school estimates shows the aggregated effect of a one-standard-deviation change in teacher quality in *each subject* in which variation in teacher quality is relevant. For reading, where interaction effects enter into the model, I show results as if the interaction effect was zero. Table 1.8 shows that the interaction effect will be negative, implying that the comparison in Table 1.12 overstates the aggregate effect of a distributional shift in secondary-school teacher quality in reading.

Table 1.12 shows, quite convincingly, that elementary-level student performance is more heavily influenced by variation in teacher quality than is secondary-level student performance.

I.XI. Conclusion

The teacher quality literature has generally assumed that off-subject teachers in secondary school do not affect student performance (e.g., math teachers do not affect reading outcomes). By relaxing this assumption, I show that educational production in secondary school is characterized by joint production among teachers

[49] I use one additional year of data here.

and that these previously-ignored, off-subject teachers can have important effects on student achievement. In math, I show that math and social studies teachers influence test scores. In reading, both English and math teachers affect performance. In both subjects and for each teacher type, distributional shifts in teacher quality can have important effects on student outcomes.

The results from this analysis are directly applicable to incentive design and teacher accountability. First, I show that the presence of joint production in secondary school should not be viewed as an obstacle to the development of performance-based incentives for teachers. In math, although student achievement is the result of inputs from multiple teachers, estimated value-added coefficients for the best and worst math teachers are largely unaffected by the degree to which joint production is modeled in the student-achievement specification. This result is replicated for English teachers in the reading model. Therefore, objections to the use of value-added modeling for the assignment of teacher accountability in secondary school based on the potential for contamination of teacher effects across subjects, at least among the highest- and lowest-ranked teachers, would be largely misguided.

The more relevant questions concerning teacher accountability in secondary school are (1) whether value-added estimates provide enough information about actual teacher quality to be useful for teacher evaluations and (2) how to determine which teachers should be evaluated based on student performance in which subjects.

In regard to the former question, although teacher value-added coefficients are quite noisy, they may still represent a reasonable improvement over the current methods that are most commonly used to determine teacher recruitment, retention and salaries.[50] Furthermore, employing teacher-fixed-effect coefficients in conjunction with other measures of teacher quality that are unlikely to have correlated measurement errors (for example, principals' evaluations that are not based on test scores) should diminish the impact of these errors and increase the visibility of the true signal of teacher quality.

As for the latter question, this paper provides insight into which teacher types affect student performance in which subjects in secondary school. To the extent that this information can be properly incorporated into teacher incentives, total educational output could be increased.[51] However, the incorporation of student achievement across multiple subjects into teacher evaluations should be carefully approached. One concern is that teachers may respond to incentives across subjects by taking focus away from important material in their own subject. Another is that cross-subject incentives, if improperly implemented, could increase free-riding opportunities.

---

[50] The weak link between student performance and the teacher qualifications upon which career decisions are generally made is well documented in the literature. See, for example, Aaronson, Barrow and Sander (2007), Angrist and Guryan (2003), Betts (1995), Betts, Zau and Rice (2003), Hanushek (1986, 1996), Kane, Rockoff and Staiger (2006), Koedel and Betts (2007).

[51] These incentives could potentially illicit more effort from teachers. Also, in the long run, providing more performance-based incentives to teachers could increase production by altering the pool of workers that select into teaching.

Finally, this analysis offers a unique opportunity to evaluate the effects of distributional shifts in teacher quality across schooling levels. Although variation in teacher quality plays a significant role in determining student outcomes at both the elementary and secondary levels, its influence is larger at the elementary level by a sizeable margin.

I.XII. ACKNOWLEDGEMENT

Chapter 1, in part, has been submitted for publication as it appears to the Review of Economics and Statistics. The dissertation author was the sole author of this paper.

Chapter 1 Figures

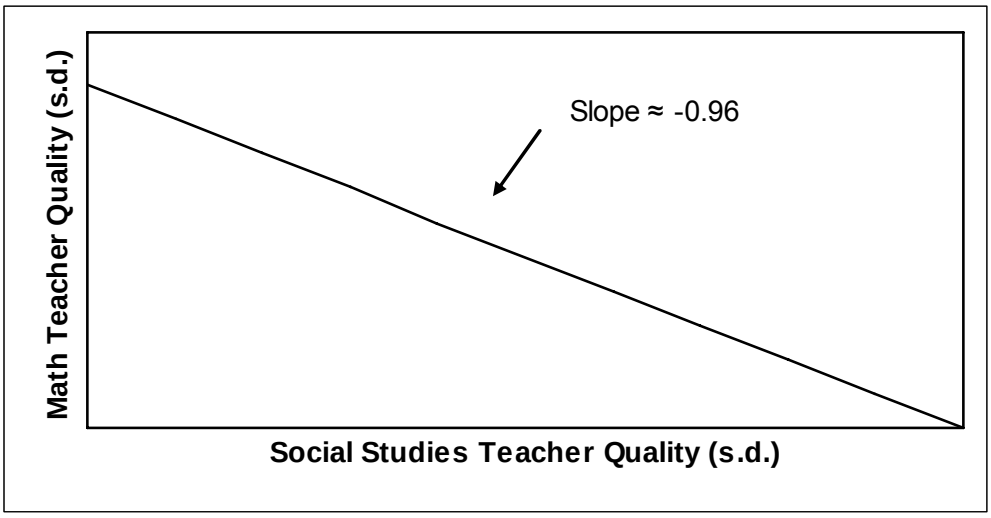

Figure 1.1. Math Production Isoquant in Math and Social Studies Teacher-Quality Space

Chapter 1 Appendix Figures

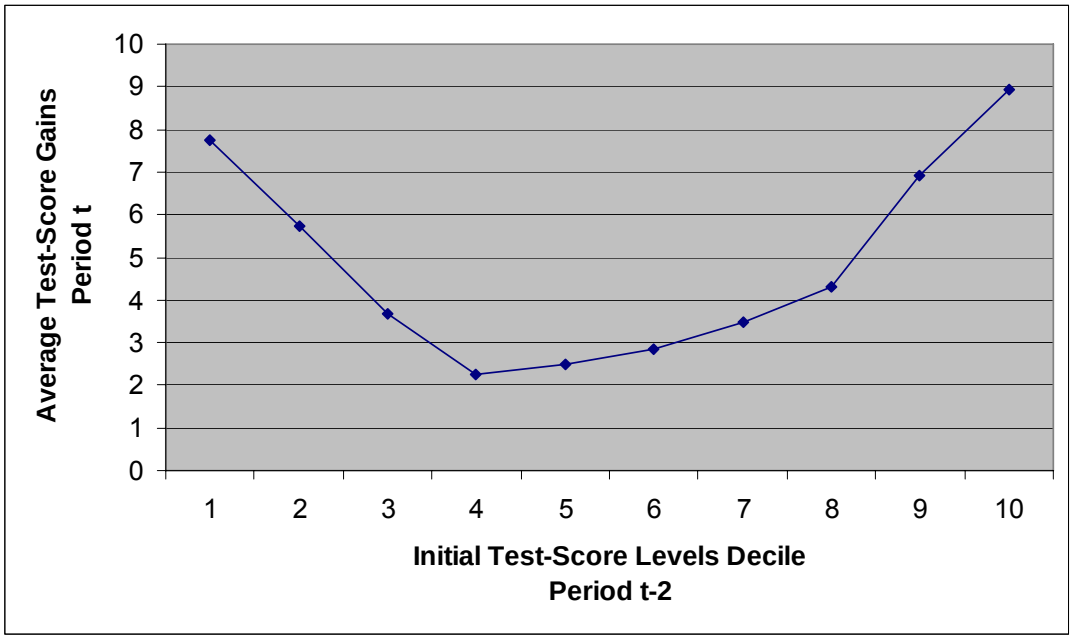

Figure 1.E.1. Achievement Gains by Decile – Math

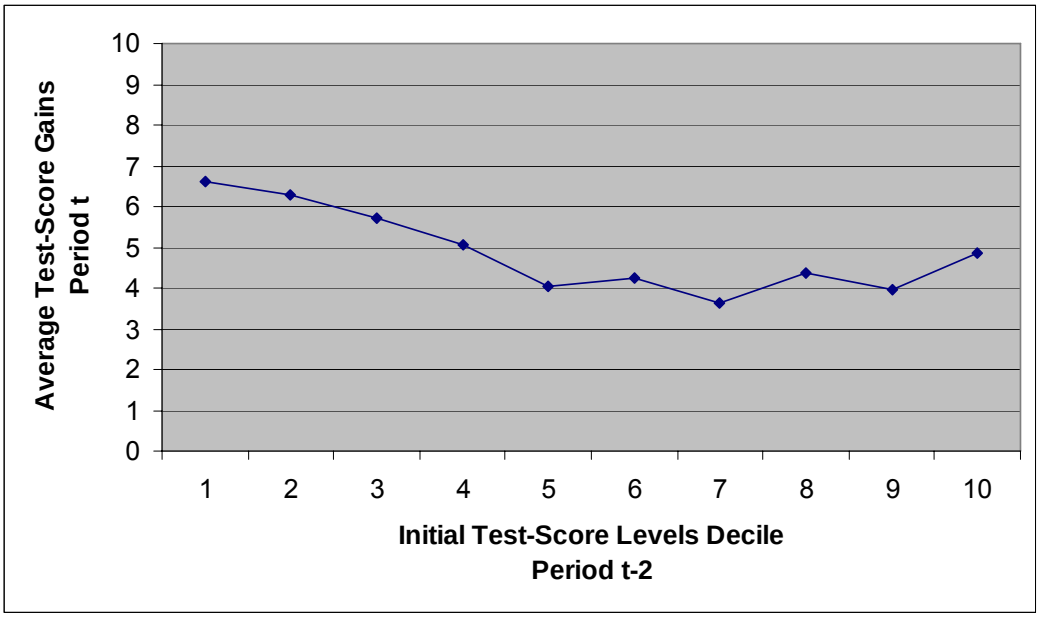

Figure 1.E.2. Achievement Gains by Decile – Reading

Chapter 1 Tables

Table 1.1. Description of Key Data Elements

Time-Varying Student Characteristics	Indicators for grade level, parental education, whether student is EL (EL = English Learner), re-designated from EL to English proficient, switched schools, accelerated a grade, held back a grade, new to the district, number of school days attended.
Time-Varying School Characteristics	Controls for the racial makeup and heterogeneity of school, school size, whether school is year round, whether school is charter or atypical, percent of school on free lunch, percent of school EL, percent of school that changed schools, percent of school new to district
Time-Varying Classroom Characteristics	Class size, peer achievement in year (t-1) - both subject-specific; subject and level of classes taken (for example, algebra or geometry, English or honors English, etc.)

Table 1.2. Class-Taking Behavior of the Student Sample by Grade Level

	Ninth Grade	Tenth Grade	Eleventh Grade
Classes Taken			
Math	100%	100%	100%
English	100%	100%	100%
Science	45%	88%	83%
Social Studies	82%	24%	99%
Science and Social Studies	27%	17%	82%

Note: Students are not tested in the twelfth grade at SDUSD.

Table 1.3. P-Values from Wald Tests for the Joint Significance of Teacher Indicator Variables in the Model of Student Achievement, by Teachers' Subject Classifications - Dependent Variable: Math Test Scores

Teachers Included by Model	Statistical Significance for Teacher Indicator Variables by Subject			
	Mathematics	English	Science	Social Studies
1. Mathematics Only	**<0.01****	-	-	-
2. Mathematics and English	0.19	0.87	-	-
3. Mathematics and Science	**<0.01****	-	0.33	-
4. Mathematics and Social Studies	**<0.01****	-	-	**<0.01****
5. Mathematics, English and Science	0.19	0.98	0.44	-
6. Mathematics, English and Social Studies	0.46	0.95	-	**0.01****
7. Mathematics, Science and Social Studies	**<0.01****	-	0.51	**<0.01****
8. Mathematics, English, Science and Social Studies	0.15	0.98	0.48	0.08

Notes: ** indicates significance with p-value ≤ 0.01

Table 1.4. P-Values from Wald Tests for the Joint Significance of Teacher Indicator Variables in the Model of Student Achievement, by Teachers' Subject Classifications - Dependent Variable: Reading Test Scores

Teachers Included by Model	Statistical Significance for Teacher Indicator Variables by Subject			
	Mathematics	English	Science	Social Studies
1. English Only	-	**<0.01****	-	-
2. English and Mathematics	**<0.01****	**<0.01****	-	-
3. English and Social Studies	-	**<0.01****	-	**<0.01****
4. English and Science	-	**<0.01****	0.27	-
5. English, Mathematics and Social Studies	**<0.01****	**<0.01****	-	**<0.01****
6. English, Mathematics and Science	**<0.01****	**<0.01****	0.34	-
7. English, Social Studies and Science	-	**<0.01****	0.59	**<0.01****
8. English, Mathematics, Social Studies and Science	**<0.01****	**<0.01****	0.27	**<0.01****

Notes: ** indicates significance with p-value ≤ 0.01

Table 1.5. Final Reading Achievement Model and Associated P-Values from Wald Tests

	Statistical Significance for Teacher Indicator Variables by Subject		
Teachers Included by Model	Mathematics	English	English-Mathematics Interactions
9. English, Mathematics and English-Mathematics Teacher Interactions	**<0.01****	**<0.01****	**<0.01****

Notes: ** indicates significance with p-value ≤ 0.01

Table 1.6. Estimated Effects of a One-Standard-Deviation Change in Teacher Quality on Student Math Achievement

	Teachers Indicator Variables Included, by Model			
	Model 1: Math Teachers Only		Model 2: Math and Social Studies Teachers	
	Unadjusted	Adjusted	Unadjusted	Adjusted
Math Teachers	0.147	0.080	0.142	0.068
Social Studies Teachers			0.110	0.065

Table 1.7. Estimated Effects of a One-Standard-Deviation Change in Teacher Quality on Student Reading Achievement

	Teachers Indicator Variables Included, by Model			
	Basic Model: English Teachers Only		Full Model: English, Math and English-Math Teacher Interactions	
	Unadjusted	Adjusted	Unadjusted	Adjusted
English Teachers	0.138	0.092	0.151	0.086
Math Teachers			0.131	0.078
English-Math Teacher Interactions			0.166	0.096

Table 1.8. Average Interaction Effects on Reading Achievement for Interactions Between English and Math Teachers by the Quintile Assignment of Each Teacher Type in their Respective Quality Distributions

		Quintile Assignments for Math Teachers				
		1	2	3	4	5 (best)
Quintile	1	0.03	0.14**	0.15**	0.01	0.05
Assignme	2	0.13**	0.10**	0.10**	-0.03	-0.02
nts for	3	0.04	0.09**	-0.05*	0.00	-0.03
English	4	0.03*	-0.01	-0.02	-0.03	-0.01
Teachers	5 (best)	-0.01	-0.10**	-0.04*	-0.23**	-0.12**

Notes: **Significant at 1% level of confidence.
*Significant at 5% level of confidence.
The results in this Table are based on 493 interactions between math and English teachers that affected at least 20 students in the dataset. The number of observations per cell ranges from 5 to 31. Estimates in just two cells are based on less than 10 observed interactions. Quintile-5 teachers are those with the highest value added, quintile-1 teachers the lowest.

Table 1.9. Estimated Correlation Coefficients Relating Teacher Fixed Effect Estimates from Restricted Models to Estimates from the Full Specification.

	(1)	(2)	(3)	(4)
Included Explanatory Variables				
(A) Lagged Test Score	Yes	Yes	Yes	Yes
(B) Student-Level Covariates	Yes	No	Yes	Yes
(C) School- and Classroom-Level Covariates, School and Subject Fixed Effects	Yes	No	No	Yes
(D) Student Fixed Effects (First Differenced)	Yes	No	No	No
Correlation Coefficient – Basic Math Model (Math Teachers Only)	1	0.26	0.28	0.63
Correlation Coefficient – Basic Reading Model (English Teachers Only)	1	0.13	0.22	0.71

Notes: Correlation coefficients compare teacher effects weighted by their standard errors. All models include indicator variables for students' grade levels. Column 1 shows the full specification to which the restricted specifications in columns 2 through 4 are compared. Wald tests reject each of the restricted models against the full model in columns 2 and 3. In columns 2 through 4, the model was estimated without first differencing. For the specifications that omit student fixed effects, additional time-invariant student-level characteristics are included (specifically, information on race and gender) and errors are clustered at the student level.

Table 1.10. Stability of Math-Teacher Value-Added Coefficients Going From the Basic to the Full Model of Student Math Achievement

		Teacher Quintile Assignments from the Full Model				
		1	2	3	4	5 (best)
Teacher	1	**87%**	9%	4%	0%	0%
Quintile	2	11%	**60%**	26%	2%	0%
Assignments	3	0%	27%	**47%**	24%	2%
from the	4	2%	2%	25%	**58%**	13%
Basic Model	5 (best)	0%	0%	0%	17%	**83%**

Note: The basic math model includes just math-teacher indicator variables, the full math model includes both math and social-studies-teacher indicator variables.

Table 1.11. Stability of English-Teacher Value-Added Coefficients Going From the Basic to the Full Model of Student Reading Achievement

		Teacher Quintile Assignments from the Full Model				
		1	2	3	4	5 (best)
Teacher	1	**78%**	20%	1%	0%	0%
Quintile	2	19%	**49%**	22%	7%	3%
Assignments	3	3%	20%	**41%**	25%	12%
from the	4	1%	9%	29%	**36%**	25%
Basic Model	5 (best)	0%	1%	6%	32%	**61%**

Note: The basic reading model includes just English-teacher indicator variables, the full reading model includes both English and math-teacher indicator variables.

Table 1.12. Effects of a One-Standard-Deviation Change in Teacher Quality (Adjusted) in Elementary and Secondary School, Measured in Standard Deviations of the Test.

	Elementary School	Secondary School (Subject Specific)	Secondary School (Aggregated)
Subject			
Math	0.26	0.07	0.13
Reading	0.19	0.09	0.16

Chapter 1 Appendix Tables

Table 1.A.1. Key Differences Between the Entire SDUSD High School Student Sample and the Final Sample Used for Estimation

	All Students	Students with 3 + Years of Data
Race		
% White	31%	30%
% Black	16%	13%
% Asian	22%	29%
% Hispanic	31%	27%
% English Learners	14%	10%
SAT 9 Math Score*	0	0.19
SAT 9 Reading Score*	0	0.20
Avg. Percentage of School on Free Lunch	44%	41%

Notes: My final sample includes 15,877 unique students with at least 3 consecutive years of test-score data out of a possible 44,846 students who could have potentially been eligible to be included based on the year that they started 9th or 10th grade. The majority of the omitted students are omitted because they do not have three contiguous years of test-score data.
*Test score performance is measured in average standard deviations from the "All Students" mean (by grade). The "all students" group includes all students at SDUSD over the entire course of the panel who had at least one completed test-score record in 9th, 10th or 11th grade.

Table 1.A.2. Key Differences Between the Entire SDUSD Teacher Sample and the Final Sample Used for Estimation – Math.

	All Teachers Who Taught at Least 50 Students in Math	Math Teachers in the Final Sample
Years Experience	10.8	14.4
% Fully Credentialed	93%	95%
% With Masters Degree	49%	53%
BA Major:		
Math	22%	54%
Education	22%	9%
Any Science	8%	7%
Social Science	18%	9%

Table 1.A.3. Key Differences Between the Entire SDUSD Teacher Sample and the Final Sample Used for Estimation - English

	All Teachers Who Taught at Least 50 Students in English	English Teachers in the Final Sample
Years Experience	11.0	13.9
% Fully Credentialed	97%	97%
% With Masters Degree	48%	52%
BA Major:		
English	37%	61%
Education	17%	5%
Social Science	21%	15%

Table 1.A.4. Key Differences Between the Entire SDUSD Teacher Sample and the Final Sample Used for Estimation - Science

	All Teachers Who Taught at Least 50 Students in Science	Science Teachers in the Final Sample
Years Experience	10.2	13.9
% Fully Credentialed	98%	97%
% With Masters Degree	49%	52%
BA Major:		
Biology	32%	48%
Chemistry	5%	12%
GeoScience	4%	6%
Physics	4%	7%
Math	3%	3%
Education	14%	5%
Social Science	13%	4%

Table 1.A.5. Key Differences Between the Entire SDUSD Teacher Sample and the Final Sample Used for Estimation – Social Studies

	All Teachers Who Taught at Least 50 Students in Social Studies	Social Studies Teachers in the Final Sample
Years Experience	12.7	13.9
% Fully Credentialed	97%	97%
% With Masters Degree	52%	55%
BA Major:		
Social Science	43%	67%
Education	20%	6%
English	11%	8%

Table 1.D.1. Sensitivity Checks for Adjusted Variance Estimates in the Basic Math and Reading Student-Achievement Specifications.

	Test-Score Levels				Value-Added			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Included Explanatory Variables								
(A) Lagged Test Score	No	No	No	No	Yes	Yes	Yes	Yes
(B) Student-Level Covariates	No	Yes	Yes	Yes	No	Yes	Yes	Yes
(C) School- and Classroom-Level Covariates, School and Subject Fixed Effects	No	No	Yes	Yes	No	No	Yes	Yes
(D) Student Fixed Effects	No	No	No	Yes	No	No	No	Yes
Adjusted Variance of Math-Teacher Quality – Basic Math Model (Standard Deviation in Parenthesis)	407.1 (20.2)	226.5 (15.1)	9.6 (3.1)	5.1 (2.3)	36.5 (6.0)	29.7 (5.4)	5.7 (2.4)	8.3 (2.9)
Adjusted Variance of English-Teacher Quality – Basic Reading Model (Standard Deviation in Parenthesis)	432.6 (20.8)	192.3 (13.9)	16.9 (4.1)	12.4 (3.5)	21.2 (4.6)	16.8 (4.1)	9.2 (3.0)	11.6 (3.4)

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.

Appendix 1.A
Data Restrictions

Section II illustrates the statistical model that seems most appropriate for accurately describing student test-score performance. This 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 requires at least three contiguous test scores per student for full identification. This data inclusion restriction reduces the available sample of students.

Additionally, I require that each student have both a math and English teacher in each year in which his or her data are used, as discussed in the text. Together, the data restrictions imposed on the student sample may bias the estimated variances of teacher quality downward by reducing student heterogeneity. Table 1.A.1 details the differences between the final sample of students used in my analysis and the general high school population at SDUSD.

As would be predicted, my final student sample is slightly advantaged relative to the SDUSD high school population as a whole. However, it is still quite diverse and generally representative of the demographics 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 1.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, I also impose participation restrictions. Kane and Staiger (2002) show that sampling variation has a significant impact on the outcomes of incentive systems based on school-level mean performance measures. Particularly, they find that schools with the smallest populations are considerably more likely to receive a reward or to be sanctioned based on student performance because the variance of the average of students' test scores from year to year is highest in these schools. A magnified version of this problem arises in my teacher analysis. In an effort to reduce the impact of sampling variation, I require that teachers have at least 20 student-years of data from my student sample to be included in the analysis.[52]

Tables 1.A.2 through 1.A.5 detail key differences between the entire SDUSD high school teacher population and the sample used in this study, by subject. In these comparisons, it was not clear how to assign the excluded teachers to a given subject. Specifically, it was unclear how many classes a teacher should have to teach in a given subject to constitute assignment to that subject. Ultimately, I included teachers into the "all teachers" sample for a given subject if, in aggregate, they taught at least 50 student-years in that subject over the course of the data panel (in this case, student years were counted for all students). This number was chosen as it corresponds to

[52] The results presented in this paper are not sensitive to a reasonable range of adjustments to this threshold.

roughly 2 class periods of students. For each of the tables below, as I increase the student-years threshold for the "all teachers" samples, these samples begin to look more and more like the final samples used in this analysis because many teachers included in the "all teachers" samples are not full-time teachers in the given subject.

Because of the imprecision in the assignment of teachers to specific subjects, Tables 1.A.2 through 1.A.5 may not reflect an "apples-to-apples" comparison. The samples used in the analysis are much more likely to reflect teachers who specialize in a specific subject. It seems intuitive that students who are taught by less specialized teachers would be subjected to more variation in teacher quality. This indicates another source of downward bias in the variance estimates presented in this paper. Unfortunately, this understatement is unavoidable given the requirements necessary to control for student fixed effects in the model of student achievement and the fact that teacher effects become less and less precisely estimated as the number of student observations per teacher falls.

Finally, note that Tables 1.A.2 through 1.A.5 may reflect some overlap. For example, if a regular science teacher taught a handful of math classes for one year due to a math-teacher shortage in that year, she would show up in the "all teachers" samples for both math and science teachers (or possibly in the "all teachers" sample for math teachers and in the "final sample" for science teachers).

Appendix 1.B
Variance Decomposition

Because the weighting matrix that I 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 1.C
Estimating an Upper Bound on the Correlation of Teacher Value-Added Across Subjects

I generate an upper bound on the correlation of math-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 the initial estimate of $corr(\hat{\theta}_m, \hat{\theta}_r)$. Consider the following:

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

The correlation coefficient of interest in this analysis is $corr(\theta_m, \theta_r)$. To obtain an upper-bound estimate, I 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 I know that $\text{cov}(\theta_m, \theta_r) \neq 0$) and expect that none of these covariance terms would be negative.[53] Given these conditions I can rewrite equation (C.1) as:

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

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

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

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

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

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

Which can once again be re-written as:

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

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. I can plug in values for the above variance components using estimates from Section III. This generates an upper bound estimate of the correlation of teacher effectiveness across subjects of approximately 1.09. Because the correlation coefficient is bounded between zero and one, we know that the correlation between the vectors of estimation errors of the math-teacher coefficients ($\underline{\lambda}_m$ and $\underline{\lambda}_r$) cannot be zero. Nonetheless, the correlation coefficient relating math-teacher quality across subjects can be bounded from above at one.

Appendix 1.D
Sensitivity Analysis

This appendix considers the sensitivity of the previously-reported estimated variances of teacher quality to the alternative specifications detailed in Table 1.9, and also to an equivalent set of specifications based on test-score levels.

The estimates reported in column 1 of Table 1.D.1 show the adjusted variance (per equation (6)) in average test-score levels, conditional on students' current grade levels, across teachers at SDUSD. These estimates incorporate not only teacher quality, but also any sorting of students and teachers throughout SDUSD across schools and classrooms. Column 2 shows that when student-level variables are included in the model, the estimated variance of the conditional teacher means declines by approximately 50 percent for both math and reading. This indicates that these variables control for a sizeable portion of the district-wide sorting that is contributing to the variance estimates in column 1. In column 3, the set of school- and classroom-level covariates and school fixed effects are added to the student-achievement specification. On the one hand, these variables control for bias in the estimated teacher effects that may result from student sorting across schools and classrooms, or from the omission of important determinants of student achievement from the model (e.g., peer quality). However, the inclusion of these controls will also reduce the estimated variance of teacher quality by removing any between-school differences in teacher performance. Finally, the estimates in column 4 incorporate

student ability measured in terms of test-score growth. The differences in the teacher-quality variance estimates between columns 3 and 4 reflect the fact that, conditional on the extensive list of controls in the model, teachers and students are positively matched within schools at SDUSD based on test-score levels.

The second vertical panel of Table 1.D.1 (columns 5 through 8) shows that the inclusion of lagged test-score performance in the student-achievement model removes much of the sorting bias (both students and teachers) in the teacher fixed effect estimates. For example, the adjusted variance of the grade-level conditional teacher means in the value-added specification is less than 10 percent of that in the levels model for both math and reading (see columns 1 and 5).

The inclusion of the student- and school-level variables in groups (B) and (C) in Table 1.D.1 have qualitatively similar effects in the value-added specification as they do in the levels specification. However, as in the levels specification, the source of the effects of these controls is unknown. It may be that they reduce omitted variables bias in the student-achievement model or they may simply remove between-school variation in teacher quality.[54] For example, the school- and classroom-level controls in variable-group (C) include measures for school-level racial composition

[54] To no avail, numerous attempts were made to disentangle the source of the effects of the sets of student-, school- and classroom-level variables. For example, school fixed effects were added separately from the other school and classroom level covariates, both before and after these covariates. The primary issue is that teachers may sort themselves into schools in ways that are aligned with the student-, school- and classroom-level covariates. Therefore, the effects of the educational inputs and environmental controls across schools cannot be separated from the effects of teacher sorting.

and peer effects. Although these measures may have a direct effect on student performance, they may also be highly correlated with teachers' preferences because teachers may prefer to work in certain socioeconomic environments and with higher-achieving students. It is possible that the best teachers are the most successful in their efforts to teach in such environments which would result in teacher sorting along these dimensions. If the differences in the adjusted-variance estimates reported in moving between columns 6 and 7 (or even 5 and 7) were entirely the result of between-school teacher sorting, estimates from column 6 (or column 5) would represent the sum of within-school and between-school variation in teacher quality.

Column 8 highlights a seemingly counterintuitive empirical result. It shows that conditional on all of the controls from the otherwise fully-specified value-added model, the inclusion of student fixed effects actually *increases* the estimated variance of teacher quality. This implies that the omission of student ability, measured by student test-score growth, is biasing the teacher coefficients from the model in column 7 *downward*. Initially, this result seems quite unlikely if students and teachers are expected to be positively matched. However, Koedel and Betts (2007) show that when the testing instrument used to measure teacher value added exhibits a test-score ceiling, negative sorting of teacher quality and student test-score performance, measured in growth, is possible.

Test-score ceiling effects are characterized by students experiencing systematic declines in test-score gains as they advance in the test-score levels distribution.[55] Importantly, these effects may be felt by students throughout the test-score distribution. In the presence of a test-score ceiling and under the assumption that students and teachers are positively matched, students assigned to the most effective teachers (the most able students) will experience restricted gains in achievement while students assigned to the least effective teachers will, relative to their high-ability peers, experience larger gains. In the absence of student fixed effects that control for students' test-score trajectories, value-added estimates for high-quality teachers will be understated due to positive student-teacher matching while for low-quality teachers, value-added estimates will be overstated. Overall, teacher value-added coefficients will be biased *toward zero* because of student sorting among teachers. In turn, this will lead to an understatement of the variance of teacher quality.[56] The test-score ceiling properties of the standardized exams used in this study are documented in Appendix 1.E. They are consistent with the findings in Table 1.D.1.

Regardless of the direction of the effect, the inclusion of student fixed effects into the student-achievement specification reduces omitted variables bias from student-teacher matches based on unobserved student characteristics. In this case, the

[55] This relationship can exist in the absence of a test-score ceiling due to regression to the mean. However, a test-score ceiling would magnify this relationship and, most importantly, limit the ability of the testing instrument to convey important information about human capital growth.
[56] For a more detailed discussion, see Koedel and Betts (2007).

direction of the bias is toward zero, indicating that estimates from the model in column 7 may understate the degree of within-school variation in teacher quality in secondary school.

Overall, the within-school and student variance estimates reported throughout this paper are from the most complete model of student achievement available. However, between-school variation in teacher quality across San Diego high schools may be non-negligible and because of this, the estimates presented in this study may understate the total variance of high school teacher quality by their omission of this between-school variance. For example, in math, if the between-school variance estimates in column 6 did not suffer from any omitted variables bias and were simply a reflection of the within-plus-between variance of math-teacher quality, estimates from the basic math model reported in Table 1.6 would understate the effect of a one-standard-deviation improvement in high school teacher quality (within plus between) by approximately 46 percent. For the basic reading model, such a scenario would imply a similar understatement of roughly 17 percent.

Appendix 1.E
Quantitative Properties of the Stanford 9 Exams in High School

This appendix details the quantitative properties of the math and reading Stanford 9 exams administered to high school students at SDUSD. Specifically, it focuses on the extent to which these exams are characterized by test-score ceilings, as test-score ceiling effects can play a significant role in the estimation of the variance of outcome-based teacher quality (see Koedel and Betts, 2007).

As indicated in Appendix 1.D, a test-score ceiling is characterized by a consistent decline in test-score gains as students make progress in the test-score levels distribution. Importantly, students need not be "at the ceiling" to be affected by it. Hanushek, Kain, O'Brien and Rivkin (2005) and Koedel and Betts (2007) discuss the importance of test-score ceiling effects in the estimation of teacher value added in great detail. The more pronounced the test-score ceiling, the more limited is the exam in terms of measuring the value added of schooling inputs.

Because of regression to the mean, it is difficult to test for pure ceiling effects by plotting test-score gains in period (t) versus test-score levels in period (t-1) because regression to the mean should ensure a negative relationship between the two regardless of whether a test-score ceiling exists. Therefore, I group all students into achievement deciles based on their raw test-score level in period (t-2). I then look to

see if the average test-score gains for students in period (t) are lower for students in higher deciles. Figures E.1 and E.2 detail these results for math and reading, respectively.

For math, the distribution of test-score gains across the test-score-levels deciles is quite odd. On the one hand, a strong test-score ceiling is implied for students in the lower achievement deciles. However, test-score gains among students in the upper achievement deciles show no indication of a ceiling and in fact; their test scores imply an effect that is the opposite of a ceiling effect. One explanation for the relationship outlined in Figure E.1 is that the Stanford 9 math exam focuses on subject material in a way that causes "average" students to be less likely to experience gains because of the classes that they happen to be taking. The model of student achievement used in this study controls for such a scenario by including a vector of subject indicators (i.e., indicators for whether a student took algebra, geometry, etc.) for each student in addition to the student fixed effects.

At first glance, the implied effect of the quantitative properties of the math portion of the Stanford 9 exam on the estimated variance of teacher quality is ambiguous. If we assume positive student-teacher matching in terms of ability (even within-subject) as is the norm, Koedel and Betts (2007) show that the relationship between test-score gains and test-score levels documented for students in the bottom deciles implies that the omission of student fixed effects in the student-achievement

model will lead to an understatement of the estimated variance of teacher quality. On the other hand, the same relationship in the upper deciles implies that the variance of teacher quality will be overstated in the absence of controls for student ability. A comparison of columns 7 and 8 of Table 1.D.1 indicates that the former effect is stronger. One explanation for this result is that the degree of student-teacher sorting is higher for students in lower achievement deciles.[57]

For reading, a relatively mild test-score ceiling is present for students in the lower deciles of the test-score levels distribution, but this ceiling disappears for students in deciles five through ten. The effects of this mild test-score ceiling can be seen in the teacher-quality variance estimates in columns 7 and 8 of Table 1.D.1.

[57] This would be the case if, for example, there is more variation in unobserved student ability among lower-achieving students or more variation in teacher quality among teachers who teach lower-achieving students.

References

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

Anderson T.W., and C. Hsiao, "Formulation and Estimation of Dynamic Models using Panel Data," *Journal of Econometrics*, 18:1 (1982), pp. 47-82.

Anderson T.W., and C. Hsiao, "Estimation of Dynamic Models with Error Components," *Journal of American Statistical Association*, 76:375 (1981), pp. 598-606.

Angrist, Joshua and Jonathan Guryan, "Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Requirements," NBER, WP 9545 (2003).

Ballou, Dale, "Do Public Schools Hire the Best Applicants," *Quarterly Journal of Economics,* 111:1 (1996), pp. 97-133.

Betts, Julian R., Andrew Zau and Lorien Rice, *Determinants of Student Achievement, New Evidence from San Diego*, Public Policy Institute of California (2003).

Betts, Julian R., "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *The Review of Economics and Statistics*, 77:2 (1995), pp. 231–250.

Hanushek, Eric, John Kain, Daniel O'Brien and Steven Rivkin, "The Market for Teacher Quality," NBER, WP 11154 (2005).

Hanushek, Eric, "Measuring Investment in Education," *The Journal of Economic Perspectives,* 10:4 (1996), pp. 9-30.

Hanushek, Eric, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature,* 24:3 (1986), pp. 1141-77.

Harris, Douglas and Tim R. Sass, "Value-Added Models and the Measurement of Teacher Quality," unpublished manuscript (2006).

Helfand, Duke, "A Formula for Failure in LA Schools," *Los Angeles Times*, January 30, 2006.

Hembree, "The Nature, Effects and Relief of Mathematics Anxiety," *Journal for Research in Mathematics Education*, 21:1 (1990), pp 33 – 46.

Ingersoll, Gary M., James P. Scamman and Wayne D. Eckerling, "Geographic Mobility and Student Achievement in an Urban Setting," *Educational Evaluation and Policy Analysis*, 11:2 (1989), pp. 143-149.

Kane, Thomas J., Jonah E. Rockoff and Douglas O. Staiger, "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," NBER, WP 12155 (2006).

Kane, Thomas and Douglas Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives,* 16:4 (2002), pp. 91-114.

Koedel, Cory and Julian Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function," University of Missouri Working Paper (2007).

McCaffrey, Daniel, J.R. Lockwood, Daniel M. Koretz and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*, RAND Corporation (2003).

Rivkin, Steven, Eric Hanushek and John Kain, "Teachers, Schools and Academic Achievement," *Econometrica*, 79:2 (2005), pp. 417-58.

Rockoff, Jonah "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review,* Papers and Proceedings, May 2004.

Rumberger, Russell W. and Katherine A. Larson, "Student Mobility and the Increased Risk of High School Dropout," *American Journal of Education*, 107:1 (1998), pp. 1 -35.

Chapter 2
Re-Examining the Role of Teacher Quality In the Educational Production Function

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.

II.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

[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).

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 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

[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, which will begin linking teacher pay to student performance in 2007.

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.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 2.1. 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

[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.

variation in teacher quality occurs within 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.

---

[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.

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 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:

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

[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.

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).

II.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 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

[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.

[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

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 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 2.1 details the controls available for students, teachers, classrooms and schools in this study. Appendix 2.A provides additional details about the data used for this project.

---

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 2.F for more details.

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

II.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 2.2 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.

Although the results in Table 2.2 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) $$Var(\hat{\theta}) = (\frac{1}{J-1})\sum_{j=1}^{J}[\hat{\theta}_j - (1/J)\sum_{j=1}^{J}(\hat{\theta}_j)]^2$$

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.

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

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

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 fixed-effects variance by the

[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 2.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.

estimation error variance on a coefficient-by-coefficient basis. See Appendix 2.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 2.3.[17]

[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 2.3 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 2.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.

The second column in Table 2.3 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 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.

The estimates of the variance of teacher quality presented in Table 2.3 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.

II.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

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

model in more detail and consider the sensitivity of our variance estimates to alternative specifications.

Table 2.4 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 2.4, 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 2.5). 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 2.4) and merits additional attention in future research.

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 student fixed effects. Table 2.5 shows eight different models from which we estimate the variance of teacher quality using the variance decomposition in equation (6).[22] The table

[21] Students do appear to be assigned to classrooms in non-random ways at SDUSD (for example, see Table 2.5 or footnote 5).

[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

indicates that the vast majority of the variation in teacher quality among elementary school teachers at SDUSD occurs within schools.

The first vertical panel of Table 2.5 (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

---

total variation in teacher quality experienced by students and because of this we do not control for teacher experience directly in our models.

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 2.5 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 2.5 suggest that the conventional wisdom that there is significant

[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.

variation in teacher value-added between schools at the elementary level may be quite inaccurate.[25]

Column 8 of Table 2.5 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 2.5 provides insight into this concern. We find that the student-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

[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.

that the bias created by this matching is *downward*. The next section explores this issue in detail.

II.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 2.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. They show that

[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.

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

[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.

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.

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

[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.

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 2.5 for the math analysis, where we find no evidence of a test-score ceiling at SDUSD (see Appendix 2.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 2.5 for the reading analysis, where we find evidence of a test-score ceiling at SDUSD (see Appendix 2.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.

[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.

II.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 $\underline{\hat{\theta}}_m$ and $\underline{\hat{\theta}}_r$ (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 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 $\hat{\underline{\theta}}_m$ and $\hat{\underline{\theta}}_r$, 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 2.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.

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

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

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 2.1).

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 2.6 reports our FGLS coefficient estimates from the weighted regression.[32,33]

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 2.6 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

[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.

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 2.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 2.1).[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 2.D for details.

[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.

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 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.

II.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

[35] An initial concern is whether teachers should be evaluated within or between schools. Because Table 2.5 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 2.8 and 2.9 below.

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 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

[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.

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.

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 2.7 demonstrates the persistence of these quintile rankings across the data subsets.

If teacher quality were perfectly observable through statistical estimation and constant over time, entries along the diagonal of Table 2.7 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 2.7 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 these teachers who would be targeted by a teacher-accountability system. Therefore, the bleak outlook portrayed in Table 2.7 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 2.7 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-

[37] Importantly, the coefficients evaluated in Table 2.7 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.

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 2.7 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 2.5) and is detailed in Table 2.8. 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 2.9.

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 2.8 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

[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.

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 2.7 and Table 2.9, 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 2.8 and 2.9 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.

II.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.

II.XI. ACKNOWLEDGEMENT

Chapter 2, in part, has been submitted for publication as it appears to the Journal of Labor Economics with Julian Betts. The dissertation author was the primary investigator and author of this paper.

Chapter 2 Tables

Table 2.1. 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

Table 2.2. 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

Table 2.3. 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

Table 2.4. 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.

Table 2.5. 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)	193.5 (13.9)	41.6 (6.5)	67.1 (8.2)	57.5 (7.6)	56.5 (7.5)	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.

Table 2.6. 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	7.35*	4.21
Authorization	(3.69)	(3.14)
Art Supplemental	2.18	4.12
Authorization	(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.

Table 2.7. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages)

Teacher Coefficient Quintile Ranking From Year t-1	Teacher Coefficient Quintile Ranking From Year t					
		1	2	3	4	5 (best)
	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.

Table 2.8. Persistence of Teacher Fixed Effects Estimates across Data Subsets (Percentages) – Between-Schools-and-Students Specification

Teacher Coefficient Quintile Ranking From Year t-1	Teacher Coefficient Quintile Ranking From Year t: 1	2	3	4	5 (best)
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 2.9. 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.

Chapter 2 Appendix Figures

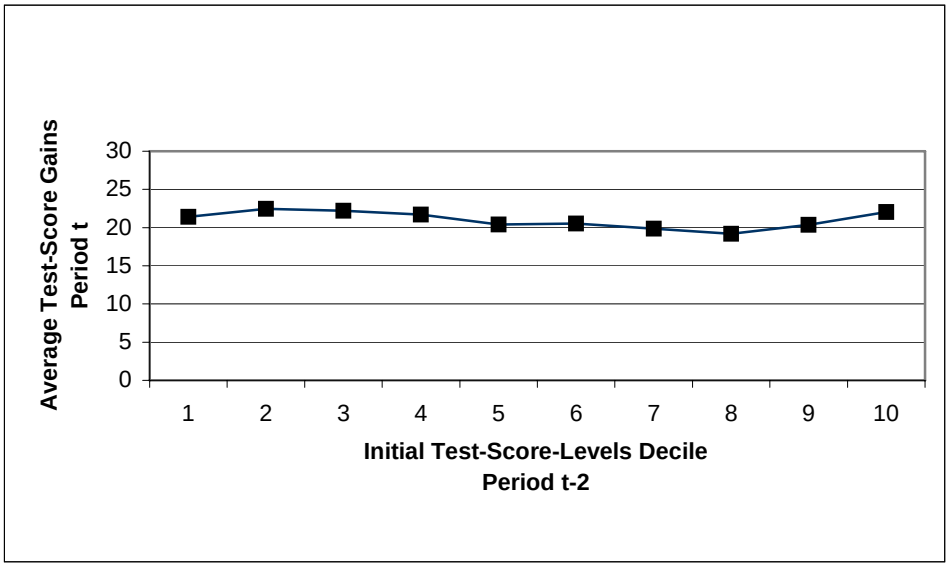

Figure 2.F.1
Achievement Gains by Decile - Math

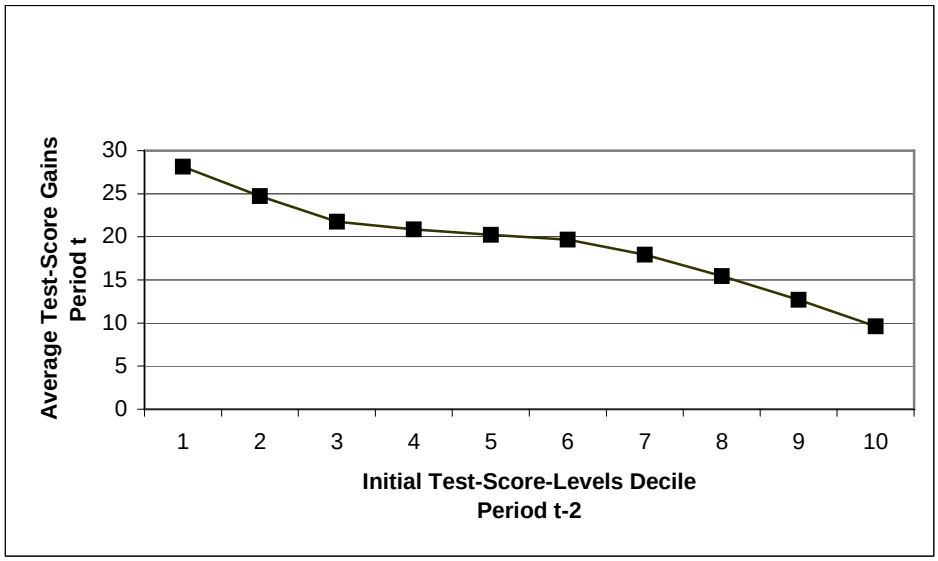

Figure 2.F.2
Achievement Gains by Decile - Reading

Chapter 2 Appendix Tables

Table 2.A.1. 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 38,369 students who would have been eligible to be included in our model based on the year that they started 2nd, 3rd, or 4th grade.
*Test score performance is measured in average standard deviations from the "All Students" mean (by grade).

Table 2.A.2. 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).

Appendix 2.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 2.A.1 details the differences between the final sample of students used in our analysis and the general elementary student population at SDUSD.

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 2.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 2.A.2 details key differences between the entire SDUSD elementary teacher population and the sample used in this study.

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 2.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:

[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.

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

This can once again be re-written as:

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

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

The R^2 statistics reported in Table 2.6 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.

[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.

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 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:

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

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:

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

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

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

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 2.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

[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.

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.

Appendix 2.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 2.F.1 and 2.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.

References

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

Anderson T.W., and C. Hsiao, "Formulation and Estimation of Dynamic Models using Panel Data," *Journal of Econometrics*, 18:1 (1982), pp. 47-82.

Anderson T.W., and C. Hsiao, "Estimation of Dynamic Models with Error Components," *Journal of American Statistical Association*, 76:375 (1981), pp. 598-606.

Angrist, Joshua and Jonathan Guryan, "Does Teacher Testing Raise Teacher Quality? Evidence From State Certification Requirements," NBER, WP 9545, 2003.

Ballou, Dale, "Do Public Schools Hire the Best Applicants," *Quarterly Journal of Economics,* 111:1 (1996), pp. 97-133.

Betts, Julian, Andrew Zau, and Lorien Rice, *Determinants of Student Achievement, New Evidence from San Diego*, Public Policy Institute of California, 2003.

Betts, Julian R., "Does school quality matter? Evidence from the National Longitudinal Survey of Youth," *The Review of Economics and Statistics,* 77:2 (1995), pp. 231-250.

Borjas, George and Glenn Sueyoshi, "A Two-Stage Estimator for Probit Models with Structural Group Effects," *Journal of Econometrics*, 64:1-2 (1994), pp.165-182.

Borjas, George, "Self-Selection and the Earnings of Immigrants," *American Economic Review*, 77:4 (1987), pp. 531-553.

Gordon, Robert, Thomas J. Kane and Douglas Staiger, *Identifying Effective Teachers Using Performance on the Job,* The Brookings Institution, 2004.

Hanushek, Eric, John Kain, Daniel O'Brien and Steven Rivkin, "The Market for Teacher Quality," NBER, WP 11154, 2005.

Hanushek, Eric, "Measuring Investment in Education," *The Journal of Economic Perspectives,* 10:4 (1996), pp. 9-30.

Hanushek, Eric, "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature,* 24:3 (1986), pp. 1141-77.

Harris, Douglas and Tim R. Sass, "Value-Added Models and the Measurement of Teacher Quality," Unpublished manuscript, Department of Economics, Florida State University, Tallahassee, 2006.

Ingersoll, Gary M., James P. Scamman and Wayne D. Eckerling, "Geographic Mobility and Student Achievement in an Urban Setting," *Educational Evaluation and Policy Analysis,* 11:2 (1989), 143-149.

Kane, Thomas E., Jonah E. Rockoff and Douglas O. Staiger, "What Does Certification Tell us about Teacher Effectiveness? Evidence from New York City," NBER, WP 12155, 2006.

Kane, Thomas and Douglas Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16:4 (2002), pp. 91-114.

Katz, Lawrence and K. Murphy, "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107:1 (1992), pp. 35-78.

Koedel, Cory, "Teacher Quality and Educational Production in Secondary School," Working Paper, University of Missouri, Columbia, 2007.

McCaffrey, Daniel, J.R. Lockwood, Daniel M. Koretz and Laura S. Hamilton, *Evaluating value-added models for teacher accountability*, RAND Corporation, 2003.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges, "How large are teacher effects?" *Educational Evaluation and Policy Analysis* 26 (2004), pp. 237-257.

Rivkin, Steven, Eric Hanushek and John Kain. "Teachers, Schools and Academic Achievement," *Econometrica* 73:2 (2005), pp. 417-458.

Rockoff, Jonah "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, Papers and Proceedings, May 2004.

Rumberger, Russell W. and Katherine A. Larson, "Student Mobility and the Increased Risk of High School Dropout," *American Journal of Education,* 107:1 (1998), pp. 1-35.

Chapter 3
Teacher Quality and Dropout Outcomes in a Large, Urban School District

Recent research shows that variation in teacher quality has large effects on student performance. However, this research is based entirely on student test scores. This paper evaluates teacher quality in terms of another educational outcome of great interest – graduation. Using a unique instrumental variables approach to identify teacher effects, I find that differences in teacher quality have large effects on graduation outcomes. Because teacher effects on graduation outcomes will be more pronounced for students who are on the graduation margin, the results imply an avenue through which high-quality teachers are more productive with disadvantaged students.

I would like to thank Andrew Zau and many administrators at San Diego Unified School District, in particular Karen Bachofer and Peter Bell, for assistance with data issues. I also thank Julian Betts, Julie Cullen, Yixiao Sun, Nora Gordon, Daniel Millimet and participants at the UCSD applied seminar for useful comments and suggestions and 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 and directed by Julian Betts.

III.I. Introduction

Recent research shows that teacher quality has large effects on student performance throughout the schooling process.[1] However, this research relies exclusively on student test scores to measure teacher quality. This paper contributes to the literature by evaluating teachers in terms of another educational outcome of great interest – graduation. The empirical results indicate that, like test scores, graduation outcomes can be heavily influenced by teacher quality.

Education Secretary Margaret Spellings recently referred to a small group of largely urban schools as "dropout factories" and did so with good reason – these schools are graduating less than 50 percent of their students. In fact, Orfield et al. (2004) show that in almost half of the high schools in the 100 largest urban school districts in the country, 12th-grade classes are less than 50 percent of the size of 9th-grade classes four years earlier. This "graduation rate crisis," as it has been called by these authors, is of great economic significance. For example, Ashenfelter and Krueger (1994) estimate that an extra year of schooling corresponds to a 12 to 16 percent increase in wages and Barrow and Rouse (2004) estimate that in 2003, high school graduates earned approximately 75 percent more than high school dropouts annually.[2] Furthermore, in addition to the costs of dropping out borne by individuals,

[1] See, for example, Koedel (2007), Koedel and Betts (2007), Rivkin, Hanushek and Kain (2006), Hanushek, Kain, O'Brien and Rivkin (2005), Aaronson, Barrow and Sander (2007), Rockoff (2004), Nye, Konstantopoulos and Hedges (2004).

[2] Rouse (1999) does a follow-up study based on Ashenfelter and Krueger and finds that these authors may have overstated the return to schooling and it is actually closer to 10 percent per year. Estimates

high dropout rates are also associated with negative externalities. Lochner and Moretti (2004) estimate that a 1-percentage-point increase in high school completion among men ages 20 to 60 would save the United States $1.4 billion per year by reducing costs associated with crime.

Given that graduation outcomes are of such great economic importance and that teacher quality has been shown to be a significant determinant of test scores, it seems natural to ask whether teacher quality affects graduation outcomes.[3] Econometrically, analyzing teacher quality in terms of graduation outcomes is complicated by the non-random assignment of students to teachers. In the test-score literature, panel datasets have been exploited to remove bias generated by this non-random student-teacher matching. Specifically, test-score studies have relied on lagged measures of performance and student fixed effects to remove sorting bias. However, in the analysis of graduation outcomes these methods cannot be implemented because graduation outcomes cannot be tracked over time as can test-score outcomes. At a given point in time, a student simply drops out of school or does not.

To estimate teacher effects on graduation outcomes, I rely on an exogenous set of instrumental variables based on school-level staffing changes from year to year.

---

from Barrow and Rouse incorporate the facts that high school graduates earn higher wages and work more hours.

[3] Loeb and Page (2000) find that teacher salary increases have a positive effect on graduation rates ten years later. However, they do not evaluate the extent to which variation in teacher quality, measured at the micro level, affects graduation outcomes.

These staffing changes represent changes in the exposure of students to teachers over time. I find that differences in teacher quality have non-negligible effects on graduation outcomes, even within schools. Because teacher effects on graduation outcomes will be more pronounced for students who are on the graduation margin, my results imply an avenue through which high-quality teachers are more productive with disadvantaged students. This finding is relevant in the debate over which types of students benefit more from high-quality teachers. Whereas recent research by Clotfelter, Ladd and Vigdor (2006), based on student test-score performance, indicates that the returns to teacher quality may be higher for advantaged students, the evidence here shows that it is perhaps the weakest students who have the most to gain from improvements in teacher quality.

III.II. Empirical Strategy

Teacher selection is likely to be endogenous to students' graduation outcomes. This endogeneity may manifest itself either through direct teacher selection within subjects, or through subject selection (i.e. choosing to take calculus) that affects teacher selection. I identify teacher effects using an instrumental variables approach based on changes in the exposure of students to teachers over time. Students' graduation decisions and teacher-selection decisions are jointly modeled and teacher effects are estimated via maximum likelihood.

Consider the following empirical model of the student dropout decision. Let D_i^* denote the net benefit to student *i* of dropping out of high school where:

(1) $$D_i^* = \beta_0 + X_i\beta_1 + T_i\theta_J + \varepsilon_i$$

In equation (1), the vector X_i includes controls for demographics, socioeconomic status, English-learner status, whether or not the student switched schools during high school and the initial math class taken in ninth grade for each student.[4] The J-dimensional vector T_i indicates which teachers taught student *i* during high school and is endogenous to the dropout decision. Student *i*'s decision to drop out, D_i, is a zero-one indicator that is equal to one if $D_i^* \geq 0$ and equal to zero otherwise.

Students' teacher-selection decisions throughout high school, T_i, can be similarly modeled. The vector T_i* can be interpreted as the set of net benefits to student *i* from taking a class with each teacher *j* at any point in high school.[5] For all *j* in which the corresponding entry in the vector $T_i^* \geq 0$, student *i* selects to be taught by teacher *j* in high school.[6]

---

[4] The initial-math-course controls provide a measure of pre-high school performance (because pre-high school performance determines initial math-course placement in high school) and are strong predictors of graduation. These controls are loosely analogous to lagged test-score controls in value-added models of test-score achievement. I considered including pre-high school test scores in the dropout specification but there was a substantial portion of the student sample that did not have test-score records for 8th grade. This may be largely because the analysis here focuses on underperforming schools in San Diego which tend to have the most transient student populations (see Section III).

[5] In this general framework, the *types* of classes taught by teachers may also be important (see, for example, Rose and Betts, 2004). Here, I focus entirely on the effects of math teachers.

[6] Net benefits are not calculated relative to other teacher choices. Instead, they are absolute. Theoretically, a by-year multinomial model of teacher selection may be more complete in that, in most

(2)

$$T_{i1}^{*} = \alpha_0 + X_i\alpha_1 + Z_{i1}\alpha_2 + u_{i1}$$

$$T_{i2}^{*} = \gamma_0 + X_i\gamma_1 + Z_{i2}\gamma_2 + u_{i2}$$

$$\vdots$$

$$T_{iJ}^{*} = \delta_0 + X_i\delta_1 + Z_{iJ}\delta_2 + u_{iJ}$$

Each observed T_{ij} is a dichotomous outcome equal to one if student *i* had teacher *j* at any point in high school and equal to zero otherwise. In equation set (2), Z_{ij} is a set of exogenous teacher and student-group-specific instrumental variables that are used to identify teacher effects.

The error terms ε, u_1,…, u_J are assumed to be joint-normally distributed with zero mean and variance-covariance matrix Ω:

---

cases, the choice of one teacher may exclude choosing others within years. However, the parameter space for such a model would be so large that it would be infeasible to estimate numerically.

$$\begin{matrix} \varepsilon_i \\ u_{i1} \\ \vdots \\ u_{iJ} \end{matrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \Omega\right) \qquad \textbf{(3)}$$

Where

$$\Omega = \begin{bmatrix} \sigma^2_{\varepsilon\varepsilon} & \sigma_{\varepsilon u_1} & \cdots & \sigma_{\varepsilon u_J} \\ \sigma_{\varepsilon u_1} & \sigma^2_{u_1 u_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{\varepsilon u_J} & \cdots & \cdots & \sigma^2_{u_J u_J} \end{bmatrix}$$

In (3), I assume the standard normalization $\sigma^2_{\varepsilon\varepsilon} = \sigma^2_{u_1 u_1} = \ldots = \sigma^2_{u_J u_J} = 1$.[7]

Given the equation set described in (1) and (2) and the distributional assumption in (3), the dropout decision is specified as a probit and teacher selection as a binary, endogenous determinant of graduation.

The instrument set that I use for each potential student-teacher match, Z_{ij}, uses variation in classes taught by teachers over time to capture student exposure, by cohort, to teachers. For example, consider a math teacher who teaches four classes of algebra and one class of geometry in one year. In the next year, this teacher might teach two classes of algebra and three classes of geometry. Furthermore, some teachers move in and out schools over time. Figure 3.1 shows four examples of

[7] In the absence of this assumption, the coefficients and error variances in (1) and (2) are only identified up to their proportions.

variation in the proportion of the total number of student semesters taught in different subjects over time for four different teachers used in this analysis. This variation reflects changes in the exposure of students to teachers and can be used to identify teacher effects. For example, depending on what year a given student happens to take geometry, the probability of that student being taught by teacher T_j may change simply because teacher T_j teaches more (or fewer) geometry classes in that year. Because the variation in classes taught by teachers over time is highest among math teachers, I focus exclusively on the role of math teachers in determining dropout outcomes.

To create the instrument sets for the student-teacher matches, I first create variables reflecting the shares of student semesters taught by each teacher in each school-subject-year combination.[8] These share variables include many zeros. For example, in all years in which a teacher does not teach at a given school, does not teach a specific subject, or does not teach at all, the share variables are equal to zero for that teacher.

I link students' class schedules to the shares of students taught by teachers. To motivate the instruments, assume momentarily – and incorrectly – that the courses chosen by students during high school are uncorrelated with their dropout decisions.

[8] I define seven subject types. They are pre-algebra (that is, anything below algebra), algebra, geometry, advanced geometry, intermediate algebra, advanced intermediate algebra, advanced math (pre-calculus and calculus).

If this were the case, I could create a set of subject-year indicator variables for each student at each school (i.e. a student may take the following sequence over the 4-year high school process: algebra-geometry-geometry-intermediate algebra). I could interact these with the shares of students taught by each teacher at that school in each subject-year to instrument for teacher selection. For example, the instrument set for a student who took geometry in the 1999-2000 school year at school X would include a positive interaction term for each teacher who taught geometry in the 1999-2000 school year at school X. These interaction terms would be equal to the percentage of student semesters of geometry taught by the individual teachers at school X in 1999-2000 and in this way would indicate the degree of exposure of the student to the teachers. For all other teachers, the instrument set would have "zero" entries for 1999-2000 for that student indicating that there could not possibly be a match based on the fact that the student took geometry at school X in that year.

To illustrate, consider a student, student *i*, who took geometry in 1999-2000. Also assume that this student took algebra in 1998-1999 (the year before) and assume that there were the same four math teachers at her school in both years - teachers A, B, C and D. In each year, teacher A taught 75 percent of the total semesters of algebra and Teacher B taught the remaining 25 percent. Also in each year, teachers C and D each taught 50 percent of the total semesters of geometry. In this simple example, the teacher selection equations for student *i*, including the instrument sets relevant for the 1998-1999 and 1999-2000 school years, would be:

$$T_{iA} = \alpha_0 + X_i\alpha_1 + (1)*(0.75)*\alpha_{21} + (0)*(0)*\alpha_{22} + (0)*(0.75)*\alpha_{23} + (1)*(0)*\alpha_{24} + u_{iA}$$
$$T_{iB} = \gamma_0 + X_i\gamma_1 + (1)*(0.25)*\gamma_{21} + (0)*(0)*\gamma_{22} + (0)*(0.25)*\gamma_{23} + (1)*(0)*\gamma_{24} + u_{iB}$$
$$T_{iC} = \rho_0 + X_i\rho_1 + (1)*(0)*\rho_{21} + (0)*(0.5)*\rho_{22} + (0)*(0)*\rho_{23} + (1)*(0.5)*\rho_{24} + u_{iC}$$
$$T_{iD} = \psi_0 + X_i\psi_1 + (1)*(0)*\psi_{21} + (0)*(0.5)*\psi_{22} + (0)*(0)*\psi_{23} + (1)*(0.5)*\psi_{24} + u_{iD}$$

The vector of student-level variables, X_i, is defined as in equations (1) and (2) and the instrument set for each student-teacher match consists of the probability that student *i* took either algebra or geometry in each year (in the example thus far, these probabilities are either zero or one because we are using student *i*'s actual math-course path in the instrument set) interacted with the share of student semesters taught by teacher *j* in that subject-year. For each teacher-selection equation there are four terms that comprise the instrument set, including zeros (in the equation for teacher A, these terms are assigned the coefficients $\alpha_{21} - \alpha_{24}$). The first instrument-set term is the interaction between an indicator for the student taking algebra in year 1 and the share of student semesters taught in algebra by the given teacher in year 1. The second term is the interaction between an indicator for the student taking geometry in year 1 and the share of student semesters taught in geometry by the given teacher in year 1. The third and fourth instrument-set terms are the year-2 analogs to the first and second terms. If the courses chosen by students during high school were truly uncorrelated with their dropout decisions, my instrumental variables approach would be exactly as it is shown in this example.

However, the course choices made by students in high school are not exogenous to their dropout decisions. Therefore, rather than using each student's specific math-course path, I instead use each student's entry-level math class in ninth grade to project her subsequent math-course path based on sample-wide averages. For example, for all students who took algebra in ninth grade at school X, I can map out the proportion who took each type of math class in subsequent years and in this way create an average path for all students who took algebra in ninth grade at school X. I create seven math-course paths at each school based on students' entry-level math courses.[9] For each student, I replace her endogenous math-course choices with the average math-course path that corresponds to her entry-level math course at her school, thereby removing the endogenous decisions of individual students from the instrument sets.[10] These math-course paths are not year specific because teachers who teach a given year-cohort may also affect the math-course path of that cohort. If the math-course paths were year specific, it would essentially build the effects of the treatments into the instruments.

[9] The seven math-course paths are based on the following entry-level mathematics classifications: No math, pre-algebra, pre-algebra/algebra, algebra, algebra/geometry, geometry, advanced geometry. Pre-algebra/algebra and algebra/geometry indicate split years. None of the (few) students who entered high school at a level above advanced geometry failed to graduate high school. Therefore, these students were omitted from the analysis. Recall that students' entry-level math classes are included directly into the dropout equation in addition to the teacher selection equations in the empirical model.

[10] An example of the final instrument set with the substitution for students' individual math-course choices is available in Appendix A. The final instrument sets are strong predictors of teacher selection.

III.III. Data

This study uses administrative data from the San Diego Unified School District (SDUSD) following high school students and teachers over time. Students and teachers are linked at the classroom level. SDUSD is the second largest school district in California (enrolling approximately 141,000 students in 1999-2000) and the student population is approximately 27 percent white, 37 percent Hispanic, 18 percent Asian/Pacific Islander and 16 percent black. 28 percent of the students at SDUSD are English Learners, and 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 at SDUSD trailed very slightly behind the national average in reading in 1999-2000. On the contrary, SDUSD students narrowly exceeded national norms in math.[11] The 4-year derived dropout rate at SDUSD in 1999-2000 was approximately 13 percent.[12,13]

[11] District characteristics summarized from Betts, Zau and Rice (2003).

[12] Source: California Department of Education. This rate includes dropouts from some atypical schools at SDUSD that focus on helping at-risk students. The empirical work here is based on data from just the 16 standard high schools at SDUSD. The derived dropout rate from these 16 schools is somewhat lower.

[13] The dropout outcome may be measured with some error as a result of the data collection process. Essentially, when a student leaves, the district relies on the student's new school to request that student's transcripts to verify that the student did not drop out. If such a request is not made, SDUSD will use the available contact information for the student to track him or her down and determine whether a dropout has occurred. In cases where transcripts are not requested and the student cannot be reached, the student will generally be considered a dropout. Hausman et al. (1998) show that probit estimates may be inconsistent when there is measurement error in the dependent variable. Although I also considered linear system IV models, the misspecification of the multiple linear probability models generated counterintuitive results. For example, in these linear SIV models at some schools, none of the entry-level-math-course controls were statistically significant predictors of graduation (for the analogous multivariate probit estimates, see Appendix Tables 3.A.1 – 3.A.4).

Because dropouts occur in all years of high school, it is important to observe each student from the ninth through (potentially) the twelfth grade. Therefore, the student data consist of 4 successive year-cohorts beginning with students in ninth grade in 1997-1998 and ending with students in ninth grade in 2000-2001 for each school (this latter group entered the twelfth grade in the final year of the panel, 2003-2004). To be included in the analysis, students had to be enrolled at SDUSD in the ninth grade.

SDUSD is a geographically large district with 16 standard, full-enrollment high schools. However, there is considerable variation in the dropout rate across schools. For example, 4-year derived dropout rates at the school level range from less than one percent to over 20 percent. In fact, almost two thirds of the dropouts from the 16 standard high schools at SDUSD come from just 4 schools. These across-school differences in the dropout rate make it difficult to argue that teachers at each school at SDUSD are equally concerned with dropouts. That is, teachers at low-dropout-rate schools may not view deterring dropouts as a significant part of their job whereas teachers at high-dropout-rate schools, some of whom may watch one in five students fail to graduate, are unlikely to feel the same. Because teachers across schools are faced with very different dropout environments, teacher quality measured

by dropout outcomes is likely to be a more relevant measure at high-dropout-rate schools.[14]

Because of the large differences in dropout rates across schools, I do not evaluate teacher quality in terms of dropout outcomes at each school at SDUSD. Instead, I focus on teacher effects at the four schools that account for the most dropouts in San Diego. My prior is that these schools are the ones in which teacher quality is most likely to play a role in determining dropout outcomes. I identify a given student as being a part of school X's population if at any time in her schooling career she attended school X.[15]

Finally, I was unable to estimate the effects of all of the math teachers in the data because there were numerous teachers who taught just a small portion of the student sample. When these teachers were included into the multivariate model, the likelihood function did not converge because the model was unable to identify teacher selection. Instead, I focus on the ten teachers at each school who taught the largest shares of the student population. Across the four schools, these teachers taught

[14] There is also reason to expect teacher quality to play a differential role in affecting dropout outcomes across schools on the student side. Specifically, students dropping out from low-dropout-rate schools are more likely to be extreme outliers whereas students dropping out from high-dropout-rate schools may be closer to the margin such that teacher quality may be more likely to make a difference.

[15] With regard to the instruments, the school population dynamics are built into the teacher shares. That is, student j who takes algebra at school Z in year 1 but then transfers to school Y and is part of school Y's population will be considered to be part of school Y's population in year 1 also. Consider the extreme case where student j is the only student that is in school Y's population but is not at school Y in year 1. Consider teacher B who teaches all N of the students who take algebra at school Y in year 1. Then teacher B's share of algebra students in year 1 at school Y will be N/(N+1), where the denominator reflects the total population of school Y as I've defined it and the numerator reflects the share of that population taught by teacher B.

between 43 and 62 percent of the total number of math-class semesters taken by students.[16] Within each school, the teacher effects are estimated relative to the average effect of the omitted teachers.[17]

III.IV. Results

The idea of an absolute "teacher effect" is inconsequential because every student has a teacher. Instead, the question of interest is whether *differences* in teacher quality can influence student outcomes. To answer this question in terms of dropouts, I estimate teacher effects from the dropout model in Section II for 40 teachers at the four schools evaluated in this study. As indicated above, these teacher effects are estimated relative to the average effect of the omitted teachers at each school where the omitted-teacher groups are comprised of the teachers who teach the fewest students. The extent to which the 40 estimated teacher effects differ from the within-school, average-omitted-teacher effects will determine the extent to which differences in teacher quality influence dropout outcomes. If teacher quality did *not* influence dropout outcomes, all teacher effects would be statistically indistinguishable from each other.

---

[16] At each school, I estimate the effects of all teachers who taught at least 5 percent of the total semesters in math taken by the student sample with the exception of four teachers. For these four teachers, there was insufficient variation in their classes taught to identify teacher selection. In place of these teachers at their respective schools, I added the teachers who taught the next most classes.

[17] The omitted teachers will bias the coefficients for the remaining teachers to the extent that the variation in classes taught between the sets of included and omitted teachers are correlated. Because of this, some of the individually estimated teacher coefficients may not be consistent estimates for the effects of their respective teachers. However, the primary motivation in this analysis is to determine whether there is a margin for teachers to influence students' dropout outcomes at all. Therefore, teacher effects that are biased only though the omission of other teacher effects will still provide valuable insight as long as they are not systematically biased towards zero.

My analysis focuses on variation in teacher quality within schools. Throughout the larger teacher-quality literature, separating school-level factors that influence student performance from across-school teacher sorting has proven difficult.[18] The primary implication of focusing on within-school variation in teacher quality is that my results will understate the potential magnitude of the effects of changes in teacher quality to the extent that quality varies across schools.

I estimate teacher effects from two different specifications for each school. First, I run a basic probit that ignores any endogeneity between teacher selection and dropout outcomes. Next, I run the multivariate probit described above in which I instrument for teacher selection and estimate teacher effects via simulated maximum likelihood.[19] Tables 3.1 through 3.4 detail the estimated teacher effects at the four schools.[20] Columns 1 and 2 in the tables report coefficient estimates from the basic and multivariate probit models. Column 3 reports marginal teacher effects from the multivariate model.[21] At schools 1 through 4 respectively, 20, 14.8, 13.5 and 9.9 percent of the student samples ultimately drop out of school. Coefficient estimates for the non-teacher components of the models are available in Appendix A.

---

[18] For more on this issue see Koedel (2007), Koedel and Betts (2007), Rivkin, Hanushek and Kain (2006) and Hanushek, Kain, O'Brien and Rivkin (2006).

[19] I use the mvprobit module in Stata by Cappellari and Jenkins (2003) to estimate the model. This module uses estimates from the individual univariate probit specifications as initial parameter values.

[20] Each school is modeled separately because the parameter space for the pooled model is large enough that the multivariate probit is infeasible to estimate. The cost of not pooling the data across schools is that the non-teacher explanatory variables are less precisely estimated.

[21] The standard errors for the marginal effects are approximated using the delta method where the explanatory variables are evaluated at their sample averages within each school.

Across the four schools, 13 out of the 40 estimated math-teacher coefficients (or 33 percent) are statistically different from the average effect of the omitted teachers indicating that differences in teacher quality can indeed affect dropout outcomes. Furthermore, the magnitudes of the estimated (marginal) teacher effects imply that they are economically meaningful, ranging from 4.3 percent to 13.4 percent. At school 3, where 13.5 percent of the student sample ultimately drops out of school, five teachers have marginal effects that, relative to the average effect of the omitted teachers, are of a magnitude greater than 6 percent. These estimates imply a significant margin by which teacher quality can affect dropout outcomes.

The point estimates for the teacher effects are predominantly negative. Of the 13 statistically significant teacher effects, 12 are negative. Overall, 33 out of 40 teacher effects have a negative sign. This implies that the teachers who teach the most students at these schools are generally more effective at reducing dropout rates than those that teach the least (recall that the average-omitted-teacher effects are based on the teachers who teach the fewest students at each school). There are numerous potential explanations for this. One possibility is that, although experience has been shown to be only weakly related to teacher performance measured by test-scores, experience with disadvantaged students may be important in deterring dropout outcomes.

Another possibility is that the results reflect selection. On the one hand, teachers who teach the fewest students at these low-performing schools may do so because they are not offered more classes than absolutely necessary by administrators because administrators knows they are of low quality. This would explain why the teachers who teach more students perform better than the omitted teachers. It may also be that teachers who teach the most students at these schools select into teaching at these schools precisely because they are effective at deterring dropout outcomes. If this were the case, some of these teachers may actually choose to work with the most disadvantaged students at these schools. The empirical results provide some support for this hypothesis. The changes in the coefficient estimates when moving from the endogenous specifications to the instrumental variables specifications (columns 1 and 2 in the tables above) imply that some of the teachers who are best at deterring dropout outcomes are matched with students who are more likely to drop out (for example, see teachers 3 and 4 at school 2, teachers 1, 3, 8 and 9 at school 3, etc.). This may reflect a concerted effort by these teachers (and administrators) to deter dropouts.

III.V. Conclusion

The effects of teacher quality, or any other educational resource for that matter, are difficult to evaluate because student-teacher matching is non-random. The outcome-based teacher quality literature, which until now has focused entirely on students' test scores, has relied on panel datasets that track student progress over time

to remove bias generated by student-teacher sorting. However, when analyzing the effects of teacher quality on other educational outcomes that cannot be tracked over time but are still of great importance, such as graduation outcomes, the econometric approaches employed in the test-score literature cannot be used. As an alternative to these methods, this study relies on an exogenous set of instrumental variables based on school-level staffing changes from year to year to estimate teacher effects on graduation outcomes. The results indicate that differences in teacher quality can play an important role in determining these outcomes. Furthermore, because the analysis is constrained to looking within schools, it is likely to understate the significance of teacher quality as an educational resource.

Finally, this study informs the debate over which types of students benefit more from high-quality teachers. Recent work by Clotfelter, Ladd and Vigdor (2006), based on student test-score performance, suggests that advantaged students may benefit more from high-quality teachers. However, the results here indicate an avenue through which high-quality teachers will be more productive with the weakest students.

III.VI. ACKNOWLEDGEMENT

Chapter 3, in part, is being prepared for publication. The dissertation author was the sole author of this paper.

CHAPTER 3 FIGURES

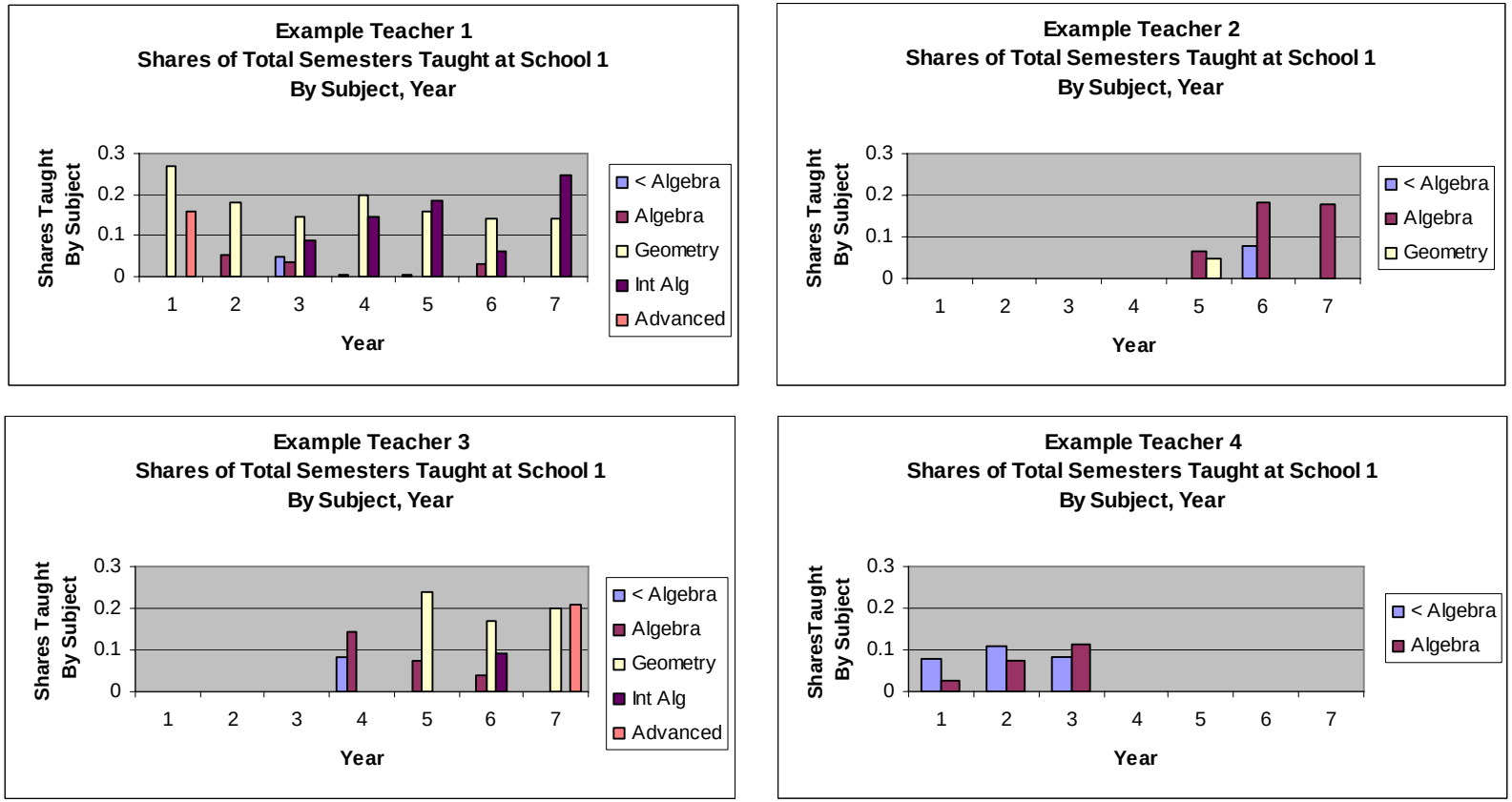

Figure 3.1. Examples of Natural Variation in Classes Taught by Subject for Four Teachers Analyzed in this Study

CHAPTER 3 TABLES

Table 3.1. Results from School 1 - Dependent Variable: Indicator for Whether a Dropout Occurred

Teacher	Basic Probit	Multivariate IV Probit	Multivariate IV Probit (Marginal Effects)
Teacher 1	-0.189	-0.101	-0.026
	(0.077)**	(0.149)	(0.041)
Teacher 2	-0.127	0.109	0.029
	(.072)*	(0.156)	(0.048)
Teacher 3	-0.045	-0.101	-0.026
	(0.076)	(0.163)	(0.047)
Teacher 4	-0.395	-0.632	-0.134
	(0.105)***	(0.182)***	(0.054)**
Teacher 5	-0.021	-0.091	-0.023
	(0.073)	(0.149)	(0.046)
Teacher 6	0.013	-0.012	-0.003
	(0.082)	(0.171)	(0.048)
Teacher 7	0.210	0.368	0.107
	(0.078)***	(0.169)**	(0.052)**
Teacher 8	-0.213	0.061	0.016
	(0.097)**	(0.169)	(0.042)
Teacher 9	0.109	-0.118	-0.030
	(0.078)	(0.166)	(0.048)
Teacher 10	-0.007	0.134	0.037
	(0.082)	(0.174)	(0.056)
Observations	3072	3072	3072

Notes: Standard errors in parentheses.
***Significant at 1% level of confidence
**Significant at 5% level of confidence
*Significant at 10% level of confidence

Table 3.2. Results from School 2 - Dependent Variable: Indicator for Whether a Dropout Occurred

Teacher	Basic Probit	Multivariate IV Probit	Multivariate IV Probit (Marginal Effects)
Teacher 1	-0.154	-0.145	-0.029
	(0.082)*	(0.139)	(0.028)
Teacher 2	-0.214	-0.045	-0.009
	(0.087)**	(0.161)	(0.032)
Teacher 3	-0.268	-0.363	-0.068
	(0.106)**	(0.162)**	(0.035)**
Teacher 4	-0.479	-0.581	-0.100
	(0.112)***	(0.175)***	(0.042)**
Teacher 5	0.164	0.169	0.037
	(0.101)	(0.173)	(0.037)
Teacher 6	-0.107	-0.236	-0.046
	(0.099)	(0.169)	(0.035)
Teacher 7	0.004	-0.124	-0.025
	(0.094)	(0.170)	(0.034)
Teacher 8	-0.087	-0.214	-0.041
	(0.108)	(0.186)	(0.038)
Teacher 9	0.116	-0.211	-0.041
	(0.112)	(0.201)	(0.041)
Teacher 10	-0.173	-0.115	-0.023
	(0.111)	(0.205)	(0.041)
Observations	2518	2518	2518

Notes: Standard errors in parentheses.
***Significant at 1% level of confidence
**Significant at 5% level of confidence
*Significant at 10% level of confidence

Table 3.3. Results from School 3 - Dependent Variable: Indicator for Whether a Dropout Occurred

Teacher	Basic Probit	Multivariate IV Probit	Multivariate IV Probit (Marginal Effects)
Teacher 1	-0.480	-0.619	-0.108
	(0.104)***	(0.158)***	(0.041)***
Teacher 2	-0.360	-0.335	-0.063
	(0.107)***	(0.180)*	(0.037)*
Teacher 3	-0.266	-0.451	-0.080
	(0.115)**	(0.185)**	(0.040)**
Teacher 4	-0.044	-0.224	-0.043
	(0.094)	(0.165)	(0.033)
Teacher 5	-0.060	-0.016	-0.003
	(0.0106)	(0.178)	(0.034)
Teacher 6	0.023	0.251	0.055
	(0.107)	(0.185)	(0.039)
Teacher 7	-0.012	0.051	0.011
	(0.122)	(0.176)	(0.035)
Teacher 8	-0.236	-0.406	-0.074
	(0.103)**	(0.162)**	(0.035)**
Teacher 9	-0.415	-0.555	-0.093
	(0.140)***	(0.185)***	(0.043)**
Teacher 10	-0.218	-0.096	-0.019
	(0.138)	(0.209)	(0.040)
Observations	2217	2217	2217

Notes: Standard errors in parentheses.
***Significant at 1% level of confidence
**Significant at 5% level of confidence
*Significant at 10% level of confidence

Table 3.4. Results from School 4 - Dependent Variable: Indicator for Whether a Dropout Occurred

Teacher	Basic Probit	Multivariate IV Probit	Multivariate IV Probit (Marginal Effects)
Teacher 1	-0.382	-0.289	-0.043
	(0.087)***	(0.140)**	(0.022)**
Teacher 2	-0.162	-0.034	-0.005
	(0.100)	(0.170)	(0.025)
Teacher 3	-0.475	-0.418	-0.057
	(0.118)***	(0.157)***	(0.027)**
Teacher 4	-0.323	-0.301	-0.043
	(0.096)***	(0.156)*	(0.023)*
Teacher 5	-0.496	-0.542	-0.069
	(0.117)***	(0.161)***	(0.028)**
Teacher 6	-0.202	-0.186	-0.028
	(0.095)**	(0.155)	(0.025)
Teacher 7	-0.064	-0.079	-0.012
	(0.088)	(0.146)	(0.024)
Teacher 8	-0.082	-0.081	-0.013
	(0.089)	(0.147)	(0.022)
Teacher 9	-0.395	-0.269	-0.038
	(0.134)***	(0.224)	(0.026)
Teacher 10	-0.194	-0.114	-0.018
	(0.118)*	(0.187)	(0.028)
Observations	3779	3779	3779

Notes: Standard errors in parentheses.
***Significant at 1% level of confidence
**Significant at 5% level of confidence
*Significant at 10% level of confidence

CHAPTER 3 APPENDIX TABLES

Table 3.A.1. Non-Teacher Results from School 1 - Dependent Variable: Indicator for Whether a Dropout Occurred

Variable	Basic Probit	Multivariate IV Probit
English Learner (EL)	0.186	0.205
	(0.063)***	(0.071)**
Re-designated from EL to non-EL	-0.606	-0.601
During High School	(0.118)***	(0.121)***
Female	-0.093	-0.085
	(0.054)*	(0.054)
Asian	0.016	0.006
	(0.149)	(0.150)
Black	-0.024	-0.038
	(0.136)	(0.136)
Hispanic	0.216	0.177
	(0.133)	(0.133)
Max Parental Ed = High School	-0.060	-0.062
	(0.126)	(0.124)
Max Parental Ed = Some College	-0.241	-0.205
	(0.156)	(0.157)
Max Parental Ed = College Graduate	-0.101	-0.133
	(0.165)	(0.166)
Max Parental Ed = Graduate School	-0.853	-0.825
	(0.464)*	(0.456)*
Max Parental Ed = Unknown	0.026	0.080
	(0.088)	(0.095)
Student changed schools mid-year at	0.166	0.171
some point during high school	(0.142)	(0.149)
9th Grade Math = No Math	-0.156	-0.144
	(0.121)	(0.126)
9th Grade Math = Part Algebra, Part	-0.159	-0.159
Pre-Algebra	(0.081)**	(0.082)*
9th Grade Math = Algebra	-0.183	-0.169
	(0.068)***	(0.073)**
9th Grade Math = Part Algebra, Part	-0.301	-0.338
Geometry	(0.178)*	(0.192)*
9th Grade Math = Geometry	-0.489	-0.393
	(0.198)**	(0.206)*
9th Grade Math = Advanced Geometry	-0.420	-0.258
	(0.210)**	(0.228)
Constant	-0.771	-0.824
	(0.154)***	(0.174)***
Observations	3072	3072

Notes: Standard errors in parentheses. All students who took intermediate algebra (> advanced geometry) in 9th grade graduated high school. Omitted variables are: Indicator variables for non-EL, non re-designated non-EL, white, parental education is high school dropout, 9th grade math class is pre-algebra and all teachers other than those listed in Table 3.1.

***Significant at 1% level of confidence.

**Significant at 5% level of confidence.

*Significant at 10% level of confidence.

Table 3.A.2. Non-Teacher Results from School 2 - Dependent Variable: Indicator for Whether a Dropout Occurred

Variable	Basic Probit	Multivariate IV Probit
English Learner (EL)	0.339	0.335
	(0.080)***	(0.083)***
Re-designated from EL to non-EL	-0.466	-0.439
During High School	(0.099)***	(0.101)***
Female	-0.037	-0.026
	(0.067)	(0.068)
Asian	-0.423	-0.453
	(0.320)	(0.317)
Black	-0.025	-0.029
	(0.148)	(0.153)
Hispanic	0.202	0.212
	(0.130)	(0.129)
Max Parental Ed = High School	0.120	0.114
	(0.153)	(0.152)
Max Parental Ed = Some College	0.088	0.049
	(0.182)	(0.181)
Max Parental Ed = College Graduate	-0.362	-0.401
	(0.224)	(0.224)*
Max Parental Ed = Graduate School	-0.007	-0.011
	(0.33)	(0.366)
Max Parental Ed = Unknown	0.037	-0.023
	(0.114)	(0.121)
Student changed schools mid-year at	0.002	-0.098
some point during high school	(0.205)	(0.218)
9th Grade Math = No Math	-0.546	0.06
	(0.330)*	(0.25)
9th Grade Math = Part Algebra, Part	-0.529	-0.509
Pre-Algebra	(0.163)***	(0.165)***
9th Grade Math = Algebra	-0.322	-0.246
	(0.082)***	(0.093)***
9th Grade Math = Geometry	-0.558	-0.496
	(0.239)**	(0.248)**
9th Grade Math = Advanced Geometry	-0.805	-0.715
	(0.149)***	(0.165)***
Constant	-0.773	-0.721
	(0.162)***	(0.179)***
Observations	2518	2518

Notes: Standard errors in parentheses. All students who took intermediate algebra (> advanced geometry) in 9th grade graduated high school. In addition, only six students took the algebra-geometry split at school 2 so that control is omitted from the model. Other omitted variables are: Indicator variables for non-EL, non re-designated non-EL, white, parental education is high school dropout, 9th grade math class is pre-algebra and all teachers other than those listed in Table 3.2.

***Significant at 1% level of confidence.

**Significant at 5% level of confidence.

*Significant at 10% level of confidence.

Table 3.A.3. Non-Teacher Results from School 3 - Dependent Variable: Indicator for Whether a Dropout Occurred

Variable	Basic Probit	Multivariate IV Probit
English Learner (EL)	0.113	0.048
	(0.085)	(0.089)
Re-designated from EL to non-EL	-0.367	-0.303
During High School	(0.120)***	(0.123)**
Female	-0.113	-0.111
	(0.073)	(0.072)
Asian	-0.036	-0.018
	(0.165)	(0.164)
Black	-0.127	-0.145
	(0.155)	(0.154)
Hispanic	0.354	0.343
	(0.155)**	(0.154)**
Max Parental Ed = High School	-0.256	-0.206
	(0.172)	(0.171)
Max Parental Ed = Some College	-0.240	-0.188
	(0.187)	(0.186)
Max Parental Ed = College Graduate	-0.056	-0.046
	(0.205)	(0.209)
Max Parental Ed = Graduate School	-0.478	-0.394
	(0.570)	(0.555)
Max Parental Ed = Unknown	-0.047	0.052
	(0.118)	(0.124)
Student changed schools mid-year at	0.397	0.342
some point during high school	(0.195)**	(0.207)*
9th Grade Math = No Math	-0.134	-0.210
	(0.147)	(0.157)
9th Grade Math = Part Algebra, Part	-0.0878	-0.099
Pre-Algebra	(0.101)	(0.105)
9th Grade Math = Algebra	-0.292	-0.369
	(0.117)**	(0.129)***
9th Grade Math = Part Algebra, Part	-0.380	-0.456
Geometry	(0.388)	(0.386)
9th Grade Math = Geometry	-0.865	-0.956
	(0.356)**	(0.358)***
9th Grade Math = Advanced Geometry	-0.981	-1.01
	(0.299)***	(0.304)***
Constant	-0.713	-0.571
	(0.187)	(0.212)
Observations	2217	2217

Notes: Standard errors in parentheses. All students who took intermediate algebra (> advanced geometry) in 9th grade graduated high school. Omitted variables are: Indicator variables for non-EL, non re-designated non-EL, white, parental education is high school dropout, 9th grade math class is pre-algebra and all teachers other than those listed in Table 3.3.

***Significant at 1% level of confidence.

**Significant at 5% level of confidence.

*Significant at 10% level of confidence.

Table 3.A.4. Non-Teacher Results from School 4 - Dependent Variable: Indicator for Whether a Dropout Occurred

Variable	Basic Probit	Multivariate IV Probit
English Learner (EL)	0.616***	0.623
	(0.088)	(0.089)***
Re-designated from EL to non-EL	-0.486	-0.488
During High School	(0.131)***	(0.132)***
Female	-0.066	-0.071
	(0.061)	(0.063)
Asian	-0.188	-0.196
	(0.129)	(0.130)
Black	0.059	0.055
	(0.133)	(0.134)
Hispanic	-0.047	-0.050
	(0.136)	(0.135)
Max Parental Ed = High School	0.065	0.075
	(0.142)	(0.142)
Max Parental Ed = Some College	-0.109	-0.110
	(0.135)	(0.136)
Max Parental Ed = College Graduate	-0.090	-0.090
	(0.133)	(0.135)
Max Parental Ed = Graduate School	0.088	0.090
	(0.285)	(0.281)
Max Parental Ed = Unknown	-0.265	-0.244
	(0.129)**	(0.130)*
Student changed schools mid-year at	-0.005	0.044
some point during high school	(0.201)	(0.200)
9th Grade Math = No Math	0.051	0.060
	(0.235)	(0.240)
9th Grade Math = Part Algebra, Part	0.149	0.132
Pre-Algebra	(0.165)	(0.167)
9th Grade Math = Algebra	-0.345	-0.354
	(0.072)***	(0.076)***
9th Grade Math = Part Algebra, Part	-0.025	-0.077
Geometry	(0.701)	(0.700)
9th Grade Math = Geometry	-0.296	-0.319
	(0.186)	(0.189)*
9th Grade Math = Advanced Geometry	-1.078	-1.067
	(0.180)***	(0.193)***
Constant	-0.588	-0.641
	(0.163)***	(0.169)***
Observations	3779	3779

Notes: Standard errors in parentheses. All students who took intermediate algebra (> advanced geometry) in 9th grade graduated high school. Omitted variables are: Indicator variables for non-EL, non re-designated non-EL, white, parental education is high school dropout, 9th grade math class is pre-algebra and all teachers other than those listed in Table 3.4.

***Significant at 1% level of confidence.

**Significant at 5% level of confidence.

*Significant at 10% level of confidence.

Appendix 3.A
Additional Details for the Dropout Analysis

For each teacher indicator variable, the set of instrumental variables used to predict student-teacher matches consists of teachers' subject-share variables interacted with the students' corresponding projected math classes, matched by year. Returning to the example in the text, I substitute for student *i*'s specific math-course path with school-wide averages given her entry-level math class. Because this student starts high school by taking algebra in ninth grade, the probability of her taking algebra in ninth grade is necessarily one. In year 2, assume that 60 percent of students who take algebra in ninth grade at school X take geometry the following year and the remaining 40 percent retake algebra. In this case, the teacher-selection equations for student *i*, including the instrument sets relevant for the 1998-1999 and 1999-2000 school years, become:

$$T_{iA} = \alpha_0 + X_i\alpha_1 + (1)*(0.75)*\alpha_{21} + (0)*(0)*\alpha_{22} + (0.4)*(0.75)*\alpha_{23} + (0.6)*(0)*\alpha_{24} + u_{iA}$$
$$T_{iB} = \gamma_0 + X_i\gamma_1 + (1)*(0.25)*\gamma_{21} + (0)*(0)*\gamma_{22} + (0.4)*(0.25)*\gamma_{23} + (0.6)*(0)*\gamma_{24} + u_{iB}$$
$$T_{iC} = \rho_0 + X_i\rho_1 + (1)*(0)*\rho_{21} + (0)*(0.5)*\rho_{22} + (0.4)*(0)*\rho_{23} + (0.6)*(0.5)*\rho_{24} + u_{iC}$$
$$T_{iD} = \psi_0 + X_i\psi_1 + (1)*(0)*\psi_{21} + (0)*(0.5)*\psi_{22} + (0.4)*(0)*\psi_{23} + (0.6)*(0.5)*\psi_{24} + u_{iD}$$

Tables 3.A.1 through 3.A.4 display the non-teacher coefficient estimates from the dropout models at schools 1 through 4, respectively.

References

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

Ashenfelter, Orley and Alan Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 84:5 (1994), pp. 1157-73.

Barrow, Lisa and Cecilia Elena Rouse, "The Economic Value of Education by Race and Ethnicity," *Economic Perspectives*, 30:2 (2006), pp. 14 – 27.

Betts, Julian R., Andrew Zau and Lorien Rice, *Determinants of Student Achievement, New Evidence from San Diego*, Public Policy Institute of California, 2003.

Cappellari, Lorenzo and Stephen P. Jenkins, "Multivariate Probit Regression Using Simulated Maximum Likelihood," *Stata Journal*, Statacorp LP, 3:3 (2003), pp. 278-294.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness, NBER, WP 11936, 2006.

Hanushek, Eric, John Kain, Daniel O'Brien and Steven Rivkin, "The Market for Teacher Quality," NBER, WP 11154, 2005.

Hausman, Jerry A., Jason Abrevaya and Fiona M. Scott-Morton, "Misclassification in the Dependent Variable in a Discrete Response Setting," *Journal of Econometrics*, 87 (1998), pp. 239–269.

Koedel, Cory, "Teacher Quality and Educational Production in Secondary School, University of Missouri Working Paper, 2007.

Koedel, Cory and Julian Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function," University of Missouri Working Paper, 2007.

Lochner, Lance and Enrico Moretti, "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests and Self-Reports," *American Economic Review*, 94:1 (2004), pp. 155 – 89.

Loeb, Susanna and Marianne E. Page, "Examining the Link between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market

Opportunities and Non-Pecuniary Variation," *The Review of Economics and Statistics*, 82:3 (2000), pp. 393-408.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges, "How large are teacher effects?" *Educational Evaluation and Policy Analysis*, 26:3 (2004), pp. 237-257.

Orfield, Gary, Daniel Losen, Johanna Wald and Christopher B. Swanson, *Losing our Future: How Minority Youth Are Being Left Behind by the Graduation Rate Crisis*, Cambridge, MA: The Civil Rights Project at Harvard University. Contributors: Advocates for Children of New York, The Civil Society Institute, 2004.

Rivkin, Steven, Eric Hanushek and John Kain, "Teachers, Schools and Academic Achievement," *Econometrica*, 79:2 (2005), pp. 417-58.

Rockoff, Jonah "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, Papers and Proceedings, May 2004.

Rose, Heather and Julian R. Betts, "The Effect of High School Courses on Earnings," *The Review of Economics and Statistics*, 86:2 (2004), pp. 497-513.

Rouse, Cecilia Elena, "Further Estimates of the Economic Return to Schooling from a New Sample of Twins," *Economics of Education Review*, 18:2 (1999), pp. 149-157.