# Teacher Quality and Joint Production in Secondary School

**Cory Koedel***
**University of Missouri**

**December 2007**

*The teacher quality literature has generally ignored teacher spillover effects in secondary school with little empirical or theoretical justification. This study uses administrative data linking students and teachers at the classroom level to show that educational output in secondary school is jointly produced by multiple teacher inputs. Specifically, math production is jointly determined by math and social studies teachers and reading production by math and English teachers. In each tested subject, distributional shifts in teacher quality for both same-subject and off-subject teachers have economically meaningful effects on student outcomes.*

*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 seminar participants at UC San Diego, UC Riverside, RAND, Southern Methodist University, Florida State University, Michigan State University and the University of Missouri 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.

Unlike elementary-school students, secondary-school students are taught by multiple teachers each year. However, the educational-production literature has traditionally assigned student performance in each subject in secondary school to a specific teacher. For example, students' math outcomes have been attributed to the effects of math teachers and reading outcomes to the effects of English teachers. The assumption that only same-subject teachers influence student performance lacks empirical support. This paper aims to quantify the degree to which teacher effects in one subject spill over into other subjects. The results show that educational output in secondary school is jointly produced by multiple teachers.

I measure teacher effects by value-added to student test scores in math and reading. In each tested subject, I consider the effects of four different teacher types: math, English, science and social studies. For math and English teachers, I estimate both same-subject and off-subject (spillover) teacher effects. Science and social studies teachers are evaluated entirely in terms of spillover effects. The primary contribution of the paper is that it explicitly models the joint-production environment in secondary school.[1] I find that math production is jointly determined by math and social studies teachers and reading production by math and English teachers. In each tested subject, distributional shifts in teacher quality for both same-subject and off-subject teachers have large effects on student performance.

By finding consistent and robust evidence for joint production in secondary education, this paper contributes to a growing literature on joint production more generally. Importantly, insights

[1] Aaronson, Barrow and Sander (2007) is the only other study that evaluates outcome-based teacher quality in secondary school. Although these authors acknowledge the possibility of joint production among secondary-school teachers, they do not pursue this issue in detail in their analysis.

from this literature are relevant to the education setting. For example, Mas and Moretti (2007) evaluate joint production among supermarket cashiers and find large peer effects among workers. These peer effects depend on the ease with which workers can observe each others' productivities. The presence of teacher spillover effects implies that teacher performance will be visible to larger teacher peer groups. School administrators may be able to exploit peer effects among secondary-school teachers to improve productivity.

Because teacher effects spill over across subjects in secondary school, analyses based on the single-teacher-effect hypothesis understate the importance of teacher quality as an educational resource. The magnitudes of the teacher spillover effects estimated here imply that this understatement is significant. Although it is not necessarily clear what the objective function of a school district is or should be, one reasonable objective function would be to maximize student achievement. Given such an objective function and numerous inputs to production, if a school district did not acknowledge teacher spillover effects it would under-allocate expenditures toward the recruitment (and possibly the development) of high-quality teachers.

Teacher spillover effects are also relevant in the context of performance-based teacher compensation. To properly design an accountability scheme, a school district would need to know the specifics of the production process as it relates to teachers. For example, which teachers should it hold accountable for student performance in which subjects? What are the relative magnitudes of the different teacher effects? Does teacher quality across subjects interact in the production process, implying the presence of team teaching?

This analysis sheds light on each of these questions. I identify which teacher types affect student performance in which subjects and provide quantitative estimates of the effects of distributional shifts in teacher quality by subject and teacher-type. I also consider the extent to which teacher effects interact in the production function. The interaction results are mixed but provide little evidence that student achievement in secondary school is team-produced. Teacher effects do interact in reading production, but not in math production. Furthermore, the interaction effects in reading production at least partly reflect diminishing returns to teacher quality across subjects rather than team-teaching effects.

## I. The Educational Production Function

The first step in evaluating joint production among teachers in secondary school is to develop a methodology for estimating teacher effects. 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:

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

Here, $Y_{isjt}$ is a test score for student *i* at school *s* with teacher-set *j* in year *t*, $\alpha_i$ represents observed and unobserved time-invariant student characteristics, $X_{it}$ is a vector of time-varying observable student characteristics, $\delta_s$ 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 $\theta_{kj}$ measures the quality of teacher k (who is part of teacher-set j). A specific form of this general value-added model, the gain-score model, is also commonly employed in empirical work.

I evaluate teacher effects on math and reading test scores for four teacher types: math, English, science and social studies.[2] 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 is defined as $(\theta_j, \theta_p, \theta_q, \theta_r)$ where $\theta_j$ indicates the quality of math teacher *j*, $\theta_p$ indicates the quality of English teacher *p*, and so on. I estimate the effects of these four teacher types on student test-score performance using the following within-school-and-student value-added specification:

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

In (1), 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. Vectors of indicator variables for schools and teachers are denoted by a "D" and are appropriately labeled. This specification allows for joint production among teachers by allowing

[2] 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 III.

multiple teachers to affect student outcomes. However, it does not allow for interactions between teachers, which will be incorporated later.

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 students take 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 student peer effects, the model includes controls for the year (t-1) achievement of classroom-level peers for each student's math and English classrooms.[3] Finally, I control for class size to prevent variation in class size from being misinterpreted as variation in teacher quality.[4]

In addition to controlling for unobserved differences in school quality, the within-school-and-student specification in (1) also minimizes omitted variables bias generated by unobserved heterogeneity in student ability across teachers. For example, if the most able students consistently sort themselves into the best teachers' classrooms (perhaps through parental lobbying), the estimated teacher effects will be unbiased by the differences in student ability across teachers created by this sorting. This is because the within-student aspect of the model ensures that teachers are evaluated relative to other teachers who teach the same students.

[3] 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.

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

The tradeoff of the within-school-and-student approach is that it ignores any between-school and between-student variation in teacher quality. To the extent that teachers vary in quality across schools or across tracks of students within schools, the within-school-and-student estimates will understate the total variance of teacher quality in secondary school. Appendix C evaluates the sensitivity of my results to alternative specifications that allow for teacher quality to vary across schools and across students within schools. The appendix provides little evidence that the within-school, across-student variance component is large.[5] However, across-school variation in teacher quality may be of a non-negligible magnitude.[6] Therefore, estimates from equation (1) may understate the variance of teacher quality in secondary school through their omission of this across-school variance. I present my results here as unbiased estimates of within-school-and-student teacher effects.

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 the model using 2SLS, instrumenting for $(TestScore^{jpqr}_{is(t-1)} - TestScore^{jpqr}_{is(t-2)})$ with $(TestScore^{jpqr}_{is(t-2)})$. The first-differenced version of equation (1) is detailed below:

[5] This finding is also supported by evidence showing that there is little within-school student tracking in the data. See below.

[6] Measuring across-school variation in teacher quality is complicated by other environmental differences across schools. In the absence of a controlled experiment, it is impossible to disentangle across-school differences in teacher quality from differences in other factors across schools that might influence student performance.

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

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.[7] The instrumentation is necessary because there is a mechanical relationship between the first-differenced lagged test score and the first-differenced error term. Namely, the period (t-1) test score is a direct function of the period (t-1) epsilon. The key assumption required for the instrumentation to be valid is that the error terms in equation (1) are serially uncorrelated (such that the period (t-2) test score is uncorrelated with the first-differenced error term). 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 epsilons and find that this primary assumption is upheld.[8,9]

## II. Identification of Teacher Effects

The identification of teacher effects is complicated by potential non-random student-teacher assignment. As discussed in the previous section, to address the more general concern that

[7] 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.

[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 $\varepsilon_{it}$'s are serially uncorrelated if the first-differenced error terms are serially correlated with a magnitude of -0.5. For students in which more than one first-differenced equation is estimated, I estimate that the serial correlation between the first-differenced error terms to be -0.45. Because I am using estimates of the first-differenced error terms to estimate this correlation, my estimate will be biased toward zero.

[9] I use robust standard errors for all 2SLS coefficients. In addition, the differenced error terms are serially correlated among 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 enforce this property of the error terms in the variance-covariance matrix for relevant students.

student ability may be correlated with teacher selection, equation (1) is first differenced. Through first differencing, this analysis focuses on within-student variation in teacher quality. However, non-random student-teacher assignment is more problematic here than in the larger literature that focuses on elementary-level teacher quality because if students are ability-tracked *across subjects*, teacher effects may be biased by other teacher effects. This section will show that in the presence of non-random student-teacher sorting, multiple teacher effects *must* be included in the model of student achievement to obtain unbiased estimates of any teacher effects.

To see this, it is perhaps most intuitive to discuss an example where it is not necessary to estimate multiple teacher effects simultaneously – when there is true random assignment of students to teachers. For illustration, consider the case where reading achievement is modeled as a function of just English teachers and we are interested in estimating the distribution of English-teacher effects. For simplicity, assume that there are only two types of teachers in secondary school - math and English – and that both types affect student performance in reading.

If students are randomly assigned into math classrooms, the average math-teacher quality experienced by any given English teacher's students will be equal to the average math-teacher quality experienced by all other English teachers' students. That is, there will be no math-teacher quality bias in the English-teacher effects (measured relative to each other). For any English teachers *p* and *p-1* who teach R and S students, respectively:

**(2)** $$(\sum_{i=r}^{R} D_{it}^{j(math)}\theta_M)/R \;=\; (\sum_{i=s}^{S} D_{it}^{j(math)}\theta_M)/S \;=\; \bar{\theta}_M$$

If condition (2) holds, unbiased English-teacher effects can be estimated from a simple model that omits controls for math teachers.

However, if students are not randomly assigned into math classrooms, the average math-teacher quality experienced by students in one English teacher's classroom need not equal the average math-teacher quality experienced by students in another's:

**(3)** $$(\sum_{i=r}^{R} D_{it}^{j(math)}\theta_M)/R \neq (\sum_{i=s}^{S} D_{it}^{j(math)}\theta_M)/S$$

In going from the case in (2) to the case in (3), math-teacher quality must be controlled for to accurately estimate English-teacher effects. That is, English teacher effects must be estimated *conditional* on math-teacher quality.

Therefore, given non-random assignment, teacher effects estimated from single-teacher-effect models of secondary-level student performance will be potentially biased by other teacher effects. One solution to remove this bias is to explicitly model *all* teacher effects, which is the approach taken here.

Although it may be necessary to estimate multiple teacher effects, one concern is that strict tracking of students to teachers may prevent these multiple teacher effects from being identified. As an example, consider the simple case illustrated in Table 2 where four students are assigned to two different English teachers and four different math teachers.

Defining each teacher in the table as a separate track, Students 1 and 2 are on different English tracks than students 3 and 4. Similarly, all 4 students are on different math tracks. One could replicate the student types from the table to create an entire population of students that is ability grouped into the classrooms of these teachers. Alternatively, one could replicate multiple "closed-loop" teacher-sharing relationships like the one illustrated in the table. In either case, teacher effects will not be fully identified because the teacher-indicator matrix will be multi-collinear.

Note, however, that the teacher effects in Table 2 are partially identified. For example, we can compare the effects of math teachers M1 and M2 to each other because students 1 and 2 share an English teacher. Similarly, we can also compare the effects of math teachers M3 and M4 to each other. However, we cannot compare the effects of math teachers M1 and M2 to the effects of math teachers M3 and M4 because such a comparison would be biased by any differences in teacher quality between English teachers (the effect of English teacher E1 will be assigned to math teachers M1 and M2 and the effect of English teacher E2 will be assigned to math teachers M3 and M4). Also, we cannot compare English teachers E1 and E2 to each other because each teacher's effect is confounded with non-overlapping math teacher effects.

In order to identify all teacher effects, at least one student must cross the tracks. Consider the addition of a $5^{th}$ student to the scenario from Table 2.

***INSERT TABLE 3***

With the addition of this 5th student, the teacher-indicator matrix is no longer multi-collinear and all teacher effects are fully identified. The effect of English teacher E2 can be estimated relative to English teacher E1 by comparing the test scores of students 2 and 5. This comparison will tell us which teacher is better and by how much. Trivially, we can use this to compute the variation in teacher quality among English teachers. As was the case in Table 2 previously, the relative effects of math teachers M1 and M2 can be identified using the test scores for students 1 and 2. Similarly, the relative effects of math teachers M3 and M4 can be identified from the test scores of students 3 and 4. Furthermore, because we know the relative effects for English teachers E1 and E2 (from students 2 and 5); the effects for math teachers M1 through M4 are comparable. Therefore, we can also estimate the variance of math-teacher quality free from any English-teacher-quality bias. Notice that all that was required to go from an unidentified model to a fully identified model is that a single student crossed the lines of the strict tracking.

Thus, the minimum requirement for the identification of multiple teacher effects is that at least one student crosses over and connects each "track" of students such that the teacher-indicator matrix is not multi-collinear. In the extreme case where identification truly relies on a single student crossing tracks, teacher effects will only be *weakly* identified. To the extent that students are heavily mixed across teachers in different subjects, approaching random assignment, teacher effects will be strongly identified.

This study is based on administrative data from the San Diego Unified School District (SDUSD). The empirical evidence on student dispersion at SDUSD, presented in the next section and in Appendix B, suggests that students are widely dispersed across teachers and that, although there

appears to be some student sorting, it is quite mild. The level of student sorting at SDUSD is such that the modeling of multiple teacher effects is both necessary and possible.

## III. Data

This study uses matched panel data from the San Diego Unified School District following high school students and teachers over time. SDUSD is the second largest school district in California (enrolling over 140,000 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. Twenty-eight 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 (Betts, Zau and Rice, 2003).

The test-score data are from the Stanford 9 test, a vertically scaled exam, and span the school years from 1997-98 through 2001-02. [10] San Diego does not attach high stakes for teachers to test-score performance; however, school-level performance is posted online and available to the public. Students at SDUSD are tested from the eighth through the eleventh grades and the data include an extensive list of school, student and classroom characteristics, which is shown in Table 1.[11]

[10] Because the Stanford 9 is vertically scaled, students' test scores do not need to be normalized. Nonetheless, in an omitted analysis I verify that all of my results are robust to models where test scores are normalized based on the San Diego distribution. Appendix E provides details on the quantitative properties of the math and reading exams.
[11] Eighth-grade test-scores are used only as (t-2) explanatory variables in the final models.

There are 16 standard high schools at SDUSD and a handful of other schools that offer secondary-level instruction (either charter schools or schools that have an atypical grade structure - for example, grades 7 – 12 or K – 9). Among the 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 primarily from students attending the standard high schools at SDUSD. However, some students from atypical or charter schools are also included.[12]

The modeling structure in equation (1) requires that all students 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 the analysis. 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 by ensuring that they are evaluated using the same student set.[13] Appendix A provides summary statistics showing that the final student sample is slightly advantaged relative to the entire student population at SDUSD but is generally representative. In Appendix C, I show that the omission of student fixed effects from the student-achievement specification results in inaccurate estimates of teacher fixed effects, justifying the empirical approach.

[12] Data from all charter and atypical schools were not available for this study. The model includes school fixed effects to control for heterogeneity in school types.

[13] 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. By grade level, Table 4 details the class-taking behavior of the student sample.

For teachers, I expect sampling variation to have a significant impact on estimated teacher effects by analogy to Kane and Staiger's analysis of school quality (2002). Thus, I require teachers to have at least 20 student-years of data to be included in the analysis.[14] Appendix A also provides summary statistics for the teacher sample.

Despite the restrictions imposed on the dataset, it still includes over 1000 teachers and more than 53,000 test-score records from over 15,000 different students.[15] Because my final samples of students and teachers are likely to be more homogeneous than their respective populations given the data inclusion restrictions, my results may understate the variance of teacher quality in secondary school.

With regard to student sorting, or ability grouping, I use two methods to evaluate the extent of its presence at SDUSD. First, I compare the average realized within-teacher standard deviations of students' period (t-1) test scores to analogous measures based on simulated student-teacher matches that are either randomly generated or perfectly sorted. If the average realized within-teacher standard deviations differ from the average within-teacher standard deviations estimated from the simulated random assignment, then ability grouping is present. This approach follows Aaronson, Barrow and Sander (2007).

The first panel of Table 5 compares realized within-teacher standard deviations of period (t-1) math test scores to various comparable measures based on simulated classroom assignments for

[14] 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.

[15] I estimate effects for 346 English teachers, 269 math teachers, 202 science teachers and 184 social studies teachers.

each teacher type. In the second panel, the analysis is repeated using period (t-1) reading test scores. The results are presented as ratios of the standard deviation of interest to the average within-grade standard deviation of the relevant test (weighted across grades, calculated using the San Diego data).

Table 5 shows that although students do not appear to be randomly assigned to teachers; the assignment pattern is much closer to what we would expect from random assignment than from perfect sorting. This implies that students are not strongly tracked, at least based on test scores, at SDUSD.[16]

In addition to showing that students are not strongly sorted to teachers using the within-teacher variance analysis above, I also use teacher-by-teacher Herfindahl indices to show that, generally speaking, students are widely dispersed from any given teacher in any subject. Appendix B details this analysis. Overall, the Herfindahl-index approach shows that students are widely dispersed to teachers across subjects in the data, corroborating the evidence from the within-teacher variance analysis in Table 5.

## IV. Methods

Because the analysis includes over 1000 teachers, tables displaying individual coefficient estimates for teachers would be difficult to interpret. Instead, I describe the variance of the distribution of teacher quality for each teacher type in each tested subject. First, I perform Wald

[16] This analysis will overstate dispersion because even if students were perfectly sorted based on true ability, noise in the test-score measures should create some within-teacher variance. However, given the magnitudes of the numbers in Table 5, measurement error should not influence the primary implication of the table.

tests for the joint significance of the sets of teacher fixed effects using equation (1). These tests evaluate the statistical significance of variation in teacher quality as a determinant of educational output and are of the form:

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

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

In (4), $\hat{\theta}$ 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 $\ell_J$ is a Jx1 vector of ones.[17] Under the null hypothesis, *W* is distributed $\chi^2_{(J-1)}$.

Although the Wald test is useful for determining 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:

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

Each fixed-effect coefficient is comprised of two components - one consisting of the true signal of teacher quality and the other of estimation error, $\hat{\theta}_j = \theta_j + \lambda_j$. Equation (5) overstates the

[17] 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.

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 $Cov(\theta,\lambda) = 0$.[18] This allows for the total variance of the teacher fixed effects to be decomposed as follows:

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

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:

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

Note that because the weighting matrix that I use for the Wald statistic is diagonal:

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

In (8), $\hat{\sigma}_j^2$ is the square of the standard error estimate for the effect of teacher *j*. Thus, scaling the Wald statistic by the number of teachers returns an estimate of the average ratio of the total fixed-effects variance to the error variance. The magnitude of the variance of the teacher-quality signal can be estimated by combining equations (6) and (7). For example, if the scaled Wald statistic is estimated to be *A* then the magnitude of the variance of the teacher-quality signal is estimated by:

[18] This assumption is not directly verifiable because both $\theta$ and $\lambda$ 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.

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

I use estimates from equation (9) to evaluate the effects of distributional shifts in teacher quality on student performance in each tested subject for each teacher type.

This approach to estimating the variance of teacher quality builds on the approach used by Aaronson, Barrow and Sander (2007). In fact, my approach would be identical to the approach of these authors if instead of using equation (7), I estimated the ratio of the total-fixed-effects variance to the estimation-error variance as:

**(10)** $$\frac{Var(\hat{\theta})}{Var(\lambda)} \approx \frac{(1/J) * \sum_{j=1}^{J} (\hat{\theta}_j - \bar{\theta})^2}{(1/J) * \sum_{j=1}^{J} (\hat{\sigma}_j^2)}$$

Although equation (10) may seem intuitive, notice that the error variance for the different teacher effects will not be constant. This is because there is heterogeneity in the number of student observations across teachers, which influences the precision of the estimates. With a non-constant error variance across teachers, equation (10) is no longer tied to the more flexible Wald statistic. The appeal of my approach is that my variance estimates are directly tied to the Wald statistic through equation (7). That is, my variance estimates are heteroskedasticity-robust.

## V. Results

*The Statistical Significance of Teacher Effects*

The results from the Wald tests for the statistical significance of variation in teacher quality, by teacher type, indicate which teacher inputs affect which test-score outputs in secondary school.[19] Tables 8 presents results from these tests for the math and reading models, as specified by 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.

Table 6 shows that variation in teacher quality among same-subject teachers is a statistically significant determinant of test scores in both math and reading for all relevant specifications.[20] In the reading models, variation in math-teacher quality is also a significant determinant of performance. 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 affect student outcomes in 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 in Table 6, 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)

[19] The magnitudes of the variances of the raw math and reading test scores are very similar. The standard deviations of these 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, respectively.

[20] 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.

from the first panel of Table 6. For reading, I add interactions between English and math teachers, English and social studies teachers, and math and social studies teachers to model (5) from the second panel of Table 6. To maintain consistency with my other data inclusion restrictions, I require teacher interactions to affect at least 20 students to be estimated.

For the interactions between math and social studies teachers in the math model, I retain the null hypothesis from the Wald test for joint significance (the p-value from this test is 0.70). Similarly, for reading output, interactions between English and social studies teachers and math 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 results in social studies teachers becoming statistically insignificant.[21] Therefore, the final reading-achievement specification is *not* model (5) from the second panel of Table 6, 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 7 details this final reading-achievement specification.

*Math-Achievement Analysis*

I start by evaluating teacher effects from the math-achievement model. First, I estimate the "basic" model that ignores the possibility of joint production among teachers (model (1) in Table

[21] 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 8. These interaction effects are jointly insignificant in the math specification and the social studies teacher indicator variables retain their statistical significance.

8). Next, I evaluate teacher effects from the full math model where indicator variables for social studies teachers are also included (model (4) in Table 8).

For each teacher type and in each model, I report the unadjusted raw variance of teacher fixed effects and the adjusted variance of teacher quality as estimated by equation (9). Results are presented as the ratio of the standard deviation of the teacher quality distribution of interest to the weighted average of the within-grade standard deviations of test scores (calculated using the San Diego data, where the weights correspond to the sample size in each grade).[22] For example, in the full model, Table 8 indicates that a one-standard-deviation increase in math-teacher quality (adjusted) corresponds to a 0.068 average within-grade standard deviation improvement in student test scores.

The results from the full math model in Table 8 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), I can calculate their slope (the marginal rate of technical substitution, MRTS). For example, an equivalent 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.

[22] 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.

The tradeoff in teacher quality between math and social studies teachers is measured by *standard deviations of each teacher-type's respective quality distribution*. It does not imply 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, the estimated MRTS 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 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.[23, 24]

Regardless of whether the results in Table 8 are driven in part by differences in heterogeneity across teacher types, the implication is unchanged. Improvements in teacher quality among math and social studies teachers can have large effects on student achievement in math.[25] Aggregating the teacher effects across subjects, a one-standard-deviation improvement in teacher quality can be expected to improve student performance by 0.13 within-grade standard deviations of the

[23] 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.

[24] Another factor that may explain the results in Table 8 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 discretion in curriculum. Variation in curriculums across social studies classes will be captured by the teacher effects, perhaps rightfully so.

[25] Of course if teacher heterogeneity is a major driver of this result, alternative recruitment practices across districts could influence which teacher-types affect student performance in which subjects. However, there is no reason to expect SDUSD to be unique among school districts in its recruitment efforts.

test.[26] When compared to the effects of other educational inputs on secondary-school math output, this implies that teacher quality is likely to be the most effective policy-relevant tool at the disposal of administrators.[27] 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.

Finally, I consider the extent to which variation in outcome-based teacher quality in math is linked to observable teacher qualifications by running another regression where I omit all of the teacher indicator variables and instead include controls for math teachers' experience, credentials, education levels and whether or not each math teacher has an undergraduate degree in mathematics.[28] None of these observable teacher qualifications have statistically significant effects in the model. Furthermore, the effects implied by their point-estimates are very small.

*Reading-Achievement Analysis*

For reading, I again start by evaluating teacher effects from the basic model in which joint production among teachers is ignored and student performance is attributed solely to variation in

[26] 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 their general result that variation in teacher quality is an important determinant of student outcomes in secondary school.

[27] The body of literature that estimates the effects of observable educational inputs on student outputs is vast. See Hanushek (1986, 1996) for literature surveys.

[28] 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.

English-teacher quality. I then evaluate the complete reading achievement model as described by model (9) in Table 7.

Similarly to the math analysis, the variance estimates from the full reading model in Table 9 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 matching/cooperation, possibly teamwork, and the effect of the compounding of teacher quality across subjects (i.e., increasing or decreasing returns). Because the data do not contain direct information on teacher quality, which I measure by student outcomes, these effects are difficult to disentangle. However, by examining the interaction effects for teachers of different 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 across subjects.

To do this, I first divide the English and math teacher effects into separate, subject-specific vectors. Within each vector, teachers are ranked from 1 to P and 1 to J, respectively, based on their value-added coefficients as estimated by the full reading model. Using these rankings, I assign all teachers to quality quintiles, where quintile-5 teachers are those with the highest value-added.

Recall from Section V that in order to maintain consistency throughout the analysis, interaction effects are estimated only for pairs of teachers who share 20 or more students. There are 493 non-exclusive pairs of teachers that meet this criterion in the data panel. Of the full samples of English and math teachers, 53 and 60 percent of these teachers, respectively, are part of at least one such pair. After ranking all teachers in each subject based on value-added to identify each teacher's quintile assignment, I use just the subsample of teachers who are involved in at least one interaction for the remainder of the interaction analysis.

Ignoring the interaction effects momentarily, I use the quintile rankings to estimate the baseline effects of teacher quality on student performance, by quintile set. A quintile set is defined by the pair of quintile rankings for a set of English and math teachers (for example, the set (1,4) would indicate an English-teacher quintile ranking of "1" and a math-teacher quintile ranking of "4"). Table 10 reports *average baseline teacher effects* – the sum of the average math-teacher effect and the average English-teacher effect, ignoring interactions - for students whose teachers are from any given quintile set. The by-quintile teacher-quality effects are centered around the (3,3) quintile, which is set to zero for ease of comparison. The cell entries are presented in terms of the same weighted average of the within-grade standard deviations of the test as the results in Table 9.[29]

By structure, the entries in Table 10 must be non-decreasing moving down and to the right. The table reflects the trivial fact that when teacher quality is measured in terms of student performance, the sum of the teacher effects for teachers in higher quintile sets will be larger.

[29] Because Table 10 displays average effects, the estimates are not adjustable for estimation error as are the variance estimates in Table 9. However, if the estimation error is independent of teachers' quintile rankings, the average estimation error in each cell of Table 10 should be zero.

Next, Table 11 incorporates the interaction effects and reports, by-quintile, *average total teacher effects*. Each cell in Table 11 is calculated as the sum of two components: (1) the analogous entry from Table 10 and (2) the average interaction effect corresponding to the relevant quintile set. Table 11 is again centered around the (3,3) quintile.

Table 11 largely retains the pattern of effects from Table 10, with two noteworthy exceptions. First, Table 11 is no longer strictly increasing moving down and to the right. For example, the table appears to imply that a student with a quintile-1 English teacher is better off with a quintile-3 math teacher than a quintile-4 math teacher. Although a literal read of the table might imply as much, this unintuitive jump is more likely the result of idiosyncrasies in the interaction effects and the arbitrariness of the quintile cutoffs. For example, the jumps in the table would shift around if teachers were divided by quartiles or sextiles instead of quintiles. Also, the averages in each cell are calculated based on relatively few pairs of teachers (ranging from just 5 to 31 pairs) and thus, they can be unduly influenced by a particular interaction or set of interactions. Because of these limitations, Table 11 is more useful for evaluating general trends than for making narrow comparisons across particular quintile sets (the same is true for Table 10, and for Table 12 to come).

The second noteworthy difference between Tables 10 and 11 is that the returns to teacher quality are less uniform once the interaction effects are incorporated. The pattern of non-uniformity introduced by the interaction effects implies that the interactions are at least partially reflective of decreasing returns to teacher quality across subjects. To show this more clearly, Table 12

isolates the interaction effects. Table 12 is generated by the cell-by-cell subtraction of Table 10 from Table 11.

Looking at Table 12, evidence of decreasing returns to teacher quality across subjects emerges. For example, consider a student who is taught by a bottom-quintile teacher in each subject. At this initially-low level of quality, the interaction effects show that improvements in teacher quality have large effects above and beyond the baseline effects. Moving southeast in the table, the interaction-based returns to improvements in teacher quality generally decline. For example, for a student who is taught by $4^{th}$-quintile English and math teachers, improvements in teacher quality to the $5^{th}$ quintile in each subject will have a net effect that is less than that of an equivalent move starting from an initially-lower quality level.

Because the production of reading output involves teacher interactions, estimating the effect of improvements in teacher quality on student performance is less straightforward than in the math analysis. However, generally speaking, the estimates in Tables 9 and 11 indicate that the effect of a one-standard-deviation improvement in math- and/or English-teacher quality can have a substantial effect on student performance.

Analogously to the math analysis, I evaluate the extent to which variation in outcome-based teacher quality in reading can be explained by observable teacher qualifications by removing all of the teacher indicator variables from equation (1) and replacing them with controls for English teachers' experience, credentials, education levels and whether or not each teacher has an undergraduate degree in English. Only the coefficient on the master's degree indicator is

statistically significant and the implied effect is small.[30] 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, the reading analysis indicates very large teacher-quality effects that are virtually unrelated to observable teacher qualifications.

## VI. The Superstar Teacher Hypothesis

The analysis from the previous section shows that math teachers affect achievement in math and reading. Does this mean that some math teachers are so great that they positively affect both math and reading performance, the proverbial "superstar teacher" effect, or similarly, 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 provide support for the superstar teacher hypothesis.

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 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 for all teachers is the same, I can get an idea of the direction of the bias introduced by the measurement error.

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

Measurement error will be smaller for teachers with a greater number of student-year observations. Therefore, I compare the correlation coefficient between $\hat{\underline{\theta}}_m$ and $\hat{\underline{\theta}}_r$ for a subset of teachers who have a relatively high number of students to an analogous correlation coefficient from 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 for math teachers. The initial estimate of the correlation between $\hat{\underline{\theta}}_m$ and $\hat{\underline{\theta}}_r$, 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 D 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) supports the superstar teacher hypothesis.[31]

[31] 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).

## VII. Test Scores and Teacher Accountability in Secondary School

Value-added is of particular policy relevance in the context of teacher accountability.[32] If a school district were interested in incorporating value-added into an accountability system for secondary-school teachers, the results from this analysis are informative because they identify which teacher inputs influence which test-score outputs and provide estimates of the effects of distributional shifts in teacher quality by teacher type. Here I address an additional question related to the practical implementation of value-added as an accountability tool: how do decisions regarding which teachers to include in the models of student achievement affect teacher rankings based on value-added? The answer to this question is important because political as well as economic considerations may be involved in the design of an accountability system.

To analyze the rank-changing effects of different levels of teacher inclusion into the models of student achievement, I consider a simple accountability system in which math teachers are evaluated based on their rankings in terms of math value-added and English teachers are evaluated based on their rankings in terms of reading value-added.[33] First, for math teachers, I estimate the basic math model that assumes only math teachers affect student math performance (Table 8, 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 math model that also allows

[32] Although value-added estimation is inherently noisy, there is evidence in the literature that value-added estimates may be more useful for evaluating teacher effectiveness than the measures currently employed by most school districts. See, for example, Aaronson, Barrow and Sander (2007), Rivkin, Hanushek and Kain (2005), and Koedel and Betts (2007).

[33] I assign each teacher an overall quality ranking 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.

for social studies teachers to also affect math performance (Table 8, 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 highest value-added. Table 13 compares the stability of these quintile assignments across the different models of student achievement. Each cell entry in Table 13 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 (here, the set (1,4) for a math teacher would indicate a quintile ranking of "1" in the basic model and "4" in the full model). 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 teachers into the model did not introduce any additional noise, the diagonal entries of Table 13 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 social studies teachers 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 13 implies a relatively low omitted variables bias generated by the omission of social studies teachers in the basic math model and indicates that a simple teacher-accountability system that rewarded math teachers based on such a model should perform

relatively well. Put differently, Table 13 shows that objections to the assignment of teacher accountability in secondary 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 includes only English-teacher effects to the full model detailed in Table 9 (including English and math teachers as well as interactions between the two). The quintile stability results are displayed in Table 14. For this analysis, the correlation between the two vectors of English-teacher coefficients is 0.87.

The results in Table 14 are similar to those in Table 13. 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.

The evidence here supports previous work showing 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).[34] Value-added modeling will be most useful in an accountability system that focuses on these teachers, which is what seems most reasonable.

## VIII. Concluding Remarks

The teacher quality literature has generally ignored teacher spillover effects in secondary school with little empirical or theoretical justification. By modeling student achievement in secondary

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

school as a function of multiple teacher inputs, I show that educational output is jointly produced. Specifically, math production is jointly determined by math and social studies teachers and reading production by math and English teachers. In each tested subject, distributional shifts in teacher quality for both same-subject and off-subject teachers have economically meaningful effects on student outcomes.

The presence teacher spillover effects implies that there are additional margins by which secondary schools can benefit from policies aimed at improving teacher quality. For example, policies aimed at improving math teacher quality can improve reading performance in addition to math performance. Furthermore, in subjects where there is a general shortage of high-quality teachers (e.g., math), schools can compensate for a lack of quality in one subject by improving quality in another. Overall, the failure to account for teacher spillover effects, which are shown here to be large, can lead to a significant understatement of the value of teacher quality as an educational resource in secondary school.

The results here are applicable to incentive design and teacher accountability. A natural extension of this work would be to determine what a system of teacher accountability might look like in practice, taking into account considerations that could not be evaluated here. For example, one concern with the implementation of across-subject teacher incentives is that teachers may respond by taking focus away from important material in their primary subjects of instruction. The degree to which across-subject teacher incentives would illicit such behavioral responses is unclear. Furthermore, depending on the objective function of the school district, these behavioral responses may or may not be desirable. For example, if a school district's

objective function disproportionately favors math achievement, the district may prefer for social studies teachers to substitute into material that improves problem-solving skills, even at the expense of the traditional social-studies curriculum. Without more information about the nature of school districts' objective functions, it is impossible to evaluate the course-material tradeoffs that are likely to be associated with the implementation of across-subject teacher incentives in secondary school.

A second concern with across-subject incentives is that if they were improperly implemented, they could increase rather than decrease free-riding opportunities among teachers. Evidence from Mas and Moretti (2007) indicates that social pressure and mutual monitoring among teacher peer groups may somewhat alleviate this concern, particularly if across-subject teacher incentives are administered in relevant subjects such that teachers can easily observe each others' productivities. Additionally, if across-subject teacher incentives were to promote the formation of teams among teachers, research by Hamilton, Nickerson and Owan (2003) implies that the positive effects of teamwork could dominate any negative effects of free-riding. The analysis here implies that these issues, drawn from the more general literature on joint production and production in teams, merit attention within the context of secondary education.

# Tables

**Table 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 2. Example of Strict Student Tracking**

| | English Teachers | | Math Teachers | | | |
|---|---|---|---|---|---|---|
| Student | E1 | E2 | M1 | M2 | M3 | M4 |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 |

**Table 3. Example of Strict Student Tracking Being Broken by a Single Student**

| | English Teachers | | Math Teachers | | | |
|---|---|---|---|---|---|---|
| Student | E1 | E2 | M1 | M2 | M3 | M4 |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 1 | 0 | 0 |

**Table 4. 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 5. Average Within-Teacher Standard Deviations of Students' Period (t-1) Test Scores in Math and Reading, by Teacher Type.**

| | | Within Schools | | Across District | |
|---|---|---|---|---|---|
| | **Actual** | **Perfect Randomization** | **Perfect Sorting** | **Perfect Randomization** | **Perfect Sorting** |
| **Math Test Scores** | | | | | |
| Math Teachers | 0.76 | 0.93 | 0.16 | 0.97 | <0.01 |
| English Teachers | 0.76 | 0.92 | 0.14 | 0.96 | <0.01 |
| Science Teachers | 0.78 | 0.91 | 0.20 | 0.96 | <0.01 |
| Social Studies Teachers | 0.77 | 0.94 | 0.20 | 0.97 | <0.01 |
| **Reading Test Scores** | | | | | |
| Math Teachers | 0.77 | 0.86 | 0.15 | 0.90 | <0.01 |
| English Teachers | 0.70 | 0.85 | 0.13 | 0.89 | <0.01 |
| Science Teachers | 0.77 | 0.86 | 0.18 | 0.90 | <0.01 |
| Social Studies Teachers | 0.72 | 0.87 | 0.19 | 0.91 | <0.01 |

Note: In the "Perfect Sorting" columns, students are sorted by period (t-1) test-score levels in math. For the randomized assignments, students are assigned to teachers based on a randomly generated number from a uniform distribution. The random assignments are repeated 25 times and estimates are averaged across all random assignments and all teachers. The estimates from the simulated random assignments are very stable across simulations.

**Table 6. P-Values from Wald Tests for the Joint Significance of Teacher Indicator Variables in the Math and Reading Models of Student Achievement, by Teacher Type**

| Teachers Included by Model | Statistical Significance for Teacher Indicator Variables by Subject | | | |
|---|---|---|---|---|
| | Mathematics | English | Science | Social Studies |
| **Math Model** | | | | |
| 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 |
| **Reading Model** | | | | |
| 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 7. Final Reading Achievement Model and Associated P-Values from Wald Tests**

| Teachers Included by Model | Statistical Significance for Teacher Indicator Variables by Subject | | |
|---|---|---|---|
| | 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 8. 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 9. 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 10. Average Baseline Effects of Teacher Quality, Excluding Interaction Effects, on Student Reading Performance by the Quintile Assignments of Each Teacher Type in their Respective Quality Distributions**

| | | Quintile Assignments for Math Teachers | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Quintile | 1 | -0.37** | -0.27** | -0.25** | -0.16** | -0.14** |
| Assignments for | 2 | -0.24** | -0.14** | -0.08** | -0.02 | 0.08** |
| English Teachers | 3 | -0.16** | -0.07** | 0.00 | 0.08** | 0.16** |
| | 4 | -0.07** | 0.01 | 0.06** | 0.14** | 0.23** |
| | 5 | -0.03 | 0.08** | 0.17** | 0.27** | 0.31** |

Notes: **Significantly different from the effect in the (3,3) quintile set at the at 1% level of confidence.
*Significantly different from the effect in the (3,3) quintile set at the 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 11. Average Total Effects of Teacher Quality on Student Reading Performance by the Quintile Assignments of Each Teacher Type in their Respective Quality Distributions**

| | | Quintile Assignments for Math Teachers | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Quintile | 1 | -0.29** | -0.08** | -0.04* | -0.11** | -0.04* |
| Assignments for | 2 | -0.06** | 0.01 | 0.07** | 0.00 | 0.11** |
| English Teachers | 3 | -0.08** | 0.07** | 0.00 | 0.13** | 0.18** |
| | 4 | 0.01 | 0.05* | 0.09** | 0.16** | 0.27** |
| | 5 | 0.01 | 0.03* | 0.17** | 0.09** | 0.24** |

Notes: **Significantly different from the effect in the (3,3) quintile set at the at 1% level of confidence.
*Significantly different from the effect in the (3,3) quintile set at the 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 12. Isolated Interaction Effects by the Quintile Assignments of Each Teacher Type in their Respective Quality Distributions for the Reading Analysis**

| | | Quintile Assignments for Math Teachers | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Quintile | 1 | 0.08** | 0.19** | 0.20** | 0.06** | 0.10** |
| Assignments for | 2 | 0.18** | 0.15** | 0.15** | 0.02 | 0.03 |
| English Teachers | 3 | 0.09** | 0.14** | 0.00 | 0.05* | 0.02 |
| | 4 | 0.08** | 0.04* | 0.03 | 0.02 | 0.04* |
| | 5 | 0.04* | -0.05* | 0.01 | -0.18** | -0.07** |

Notes: **Significantly different from the effect in the (3,3) quintile set at the at 1% level of confidence.
*Significantly different from the effect in the (3,3) quintile set at the 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 13. 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 |
| Teacher Quintile | 1 | **87%** | 9% | 4% | 0% | 0% |
| Assignments from | 2 | 11% | **60%** | 26% | 2% | 0% |
| the Basic Model | 3 | 0% | 27% | **47%** | 24% | 2% |
| | 4 | 2% | 2% | 25% | **58%** | 13% |
| | 5 | 0% | 0% | 0% | 17% | **83%** |

**Table 14. 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 |
| Teacher Quintile | 1 | **78%** | 20% | 1% | 0% | 0% |
| Assignments from | 2 | 19% | **49%** | 22% | 7% | 3% |
| the Basic Model | 3 | 3% | 20% | **41%** | 25% | 12% |
| | 4 | 1% | 9% | 29% | **36%** | 25% |
| | 5 | 0% | 1% | 6% | 32% | **61%** |

# Appendix Figures

**Figure E.1**
**Achievement Gains by Decile - Math**

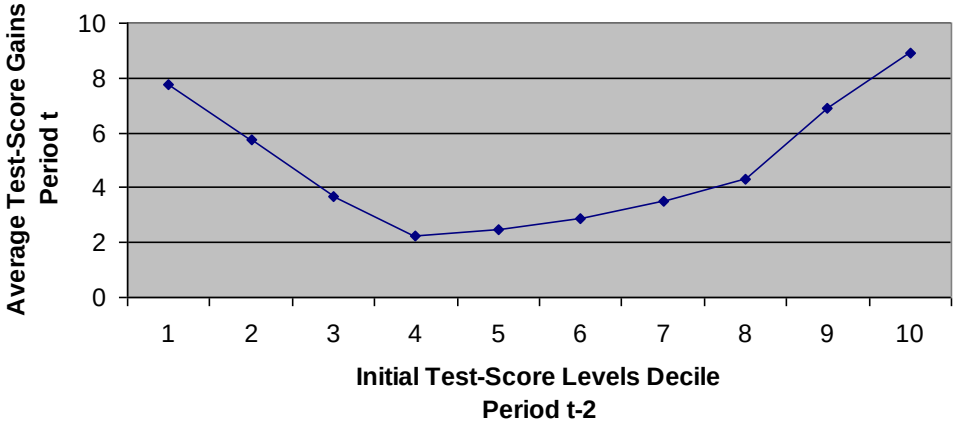
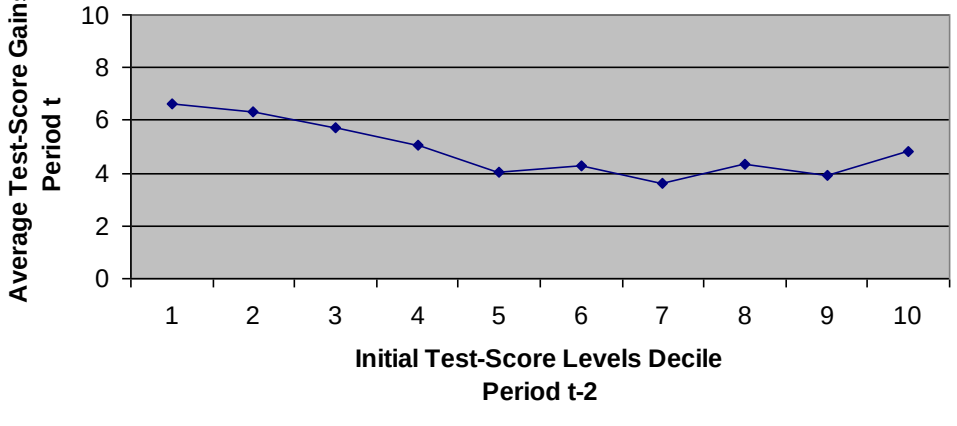



**Figure E.2**
**Achievement Gains by Decile - Reading**



# Appendix Tables

**Table 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% |

My final sample includes 15,877 unique students with at least 3 consecutive years of test-score data out of a possible 32,740 students who could have potentially been eligible to be included based on the year that they started 9$^{th}$ or 10$^{th}$ 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 9$^{th}$, 10$^{th}$ or 11$^{th}$ grade.

**Table 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 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 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 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 B.1. Average and Median Per-Teacher Herfindahl Indices for Each Teacher Type to Each Teacher Type**

| | Herfindahl Indices | |
|---|---|---|
| | Mean | Median |
| Math Teachers | | |
| To English Teachers | 0.12 | 0.11 |
| To Science Teachers | 0.10 | 0.08 |
| To Social Studies Teachers | 0.10 | 0.09 |
| English Teachers | | |
| To Math Teachers | 0.15 | 0.13 |
| To Science Teachers | 0.14 | 0.12 |
| To Social Studies Teachers | 0.16 | 0.10 |
| Science Teachers | | |
| To English Teachers | 0.18 | 0.13 |
| To Math Teachers | 0.18 | 0.14 |
| To Social Studies Teachers | 0.11 | 0.08 |
| Social Studies Teachers | | |
| To English Teachers | 0.21 | 0.16 |
| To Math Teachers | 0.17 | 0.12 |
| To Science Teachers | 0.11 | 0.08 |

Note: Unlike in math and English, students do not take science and social studies every year. To appropriately reflect the dispersion created by students without social studies and/or science teachers, each student who did not have one of these teachers was treated as going into a unique bin.

**Table B.2. Average and Median Per-Teacher Herfindahl Indices for Each Teacher Type to Each Teacher Type for a Single Year (1999-2000)**

| | Herfindahl Indices | |
|---|---|---|
| | Mean | Median |
| Math Teachers | | |
| To English Teachers | 0.16 | 0.13 |
| To Science Teachers | 0.12 | 0.10 |
| To Social Studies Teachers | 0.12 | 0.10 |
| English Teachers | | |
| To Math Teachers | 0.20 | 0.19 |
| To Science Teachers | 0.17 | 0.13 |
| To Social Studies Teachers | 0.23 | 0.18 |
| Science Teachers | | |
| To English Teachers | 0.19 | 0.17 |
| To Math Teachers | 0.18 | 0.16 |
| To Social Studies Teachers | 0.11 | 0.08 |
| Social Studies Teachers | | |
| To English Teachers | 0.27 | 0.19 |
| To Math Teachers | 0.20 | 0.18 |
| To Science Teachers | 0.14 | 0.12 |

Note: Unlike in math and English, students do not take science and social studies every year. To appropriately reflect the dispersion created by students without social studies and/or science teachers, each student who did not have one of these teachers was treated as going into a unique bin.

**Table C.1. Specification Robustness Checks for the Full Math and Reading Models**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Included Explanatory Variables | | | | |
| (A) Lagged Test Score | Yes | Yes | Yes | Yes |
| (B) Student-Level Covariates | No | Yes | Yes | Yes |
| (C) School- and Classroom-Level Covariates, School and Subject Fixed Effects | No | No | Yes | Yes |
| (D) Student Fixed Effects (First Differenced) | No | No | No | Yes |
| **Full Math Model** | | | | |
| Math Teachers | | | | |
| P-value from Wald Test for Inclusion into Model | <.01** | <.01** | <.01** | <.01** |
| Adjusted Variance Estimate | 0.158 | 0.148 | 0.065 | 0.068 |
| Correlation Coefficient | 0.31 | 0.33 | 0.64 | 1 |
| Social Studies Teachers | | | | |
| P-value from Wald Test for Inclusion into Model | <.01** | <.01** | <.01** | <.01** |
| Adjusted Variance Estimate | 0.133 | 0.123 | 0.077 | 0.065 |
| Correlation Coefficient | 0.56 | 0.56 | 0.66 | 1 |
| **Full Reading Model** | | | | |
| English Teachers | | | | |
| P-value from Wald Test for Inclusion into Model | <.01** | <.01** | <.01** | <.01** |
| Adjusted Variance Estimate | 0.140 | 0.131 | 0.108 | 0.086 |
| Correlation Coefficient | 0.35 | 0.40 | 0.61 | 1 |
| Math Teachers | | | | |
| P-value from Wald Test for Inclusion into Model | <.01** | <.01** | <.01** | <.01** |
| Adjusted Variance Estimate | 0.108 | 0.099 | 0.073 | 0.078 |
| Correlation Coefficient | 0.24 | 0.25 | 0.59 | 1 |

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

# Appendix A
# Data Restrictions

Section I 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 may bias the estimated variances of teacher quality downward by reducing student heterogeneity. Table 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 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.[35]

Tables A.2 through 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 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.

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

Because of the imprecision in the assignment of teachers to specific subjects, Tables A.2 through 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 A.2 through 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 B
# Teacher-by-Teacher Herfindahl Indices

Herfindahl indices are common in the industrial organization literature where they are used to measure industry concentration. For math teacher j who has students dispersed into the classrooms of social studies teachers r = 1,..,r, the Herfindahl index takes the following form:

$$H = \sum_{r=1}^{R} (S_{rj} / S_j)^2$$

Here, $S_{rj}$ is the share of math teacher *j*'s students taught by social studies teacher *r* and $S_j$ is the total number of students taught by math teacher *j*.

For each teacher-type, I randomly select 50 teachers and calculate each teacher's Herfindahl index into the classrooms of every other teacher type. Table B.1 presents the averages and medians of these teacher-specific Herfindahl indices. The Department of Justice considers industries where the Herfindahl index is between 0.10 and 0.18 to be moderately concentrated and industries where it is above 0.18 to be concentrated. Although the interpretation of the Herfindahl indices for teachers may be less clear, they certainly provide useful information about the concentration of teachers' students across subjects. For example, Table B.1 indicates that the average English teacher in my sample could send, at most, 36 percent of her students to any particular math teacher if she sent all of her remaining students to different math teachers. Alternatively, this average English teacher might send 15 percent of her students to six different math teachers and the remaining ten percent to a seventh math teacher.

The magnitudes of the Herfindahl indices imply that students are well-dispersed among teachers across subjects at SDUSD. There are three factors in the analysis that contribute to this dispersion. First, the high schools at SDUSD are all relatively large. Second, there are two structural factors associated with class scheduling that contribute to student dispersion into classrooms across subjects: (1) math in secondary school is not a grade-level specific subject whereas social studies, science and English generally are and (2) the typical student at SDUSD alternates between taking science and social studies in the 9th and 10th grades and, generally speaking, only takes these subjects concurrently in the 11th grade.[36] Third, the teacher effects in this analysis are not estimated by-year. Instead, unlike some other work on teacher quality, I combine all of the years of available data into my models and estimate a single teacher effect for each teacher in each tested subject. This means that teacher turnover and changes in the classes taught by teachers from year to year affect the sharing patterns of teachers across subjects.

Because the effects of year-to-year teacher turnover on the Herfindahl indices do not speak to ability grouping directly, it is also of interest to evaluate dispersion within years. Table B.2 presents average and median Herfindahl indices for a single year, 1999-2000, for 50 (newly) randomly selected teachers in each subject.[37] These indices are analogous to the full-sample indices presented in Table B.1. While the indices in Table B.1 provide information about how dispersion affects the mechanical identification of teacher effects, the indices in Table B.2 provide more specific information about how much tracking occurs across teachers in different subjects at SDUSD.

[36] See Table 4.
[37] I re-selected the teacher samples for the single-year analysis because not all teachers taught in 1999-2000.

Not surprisingly, the Herfindahl indices in Table B.2 are larger than those in Table B.1. However, Table B.2 still shows that students are widely dispersed to teachers across subjects, even within years, corroborating the evidence from the within-teacher variance analysis in Table 5 from the text.

# Appendix C
# Sensitivity Analysis

I evaluate the importance of the different components of the student-achievement specification by examining the robustness of my results to alternative models. Table C.1 shows four separate value-added specifications. The fourth column of the table shows the full model estimated in equation (1) and columns 1 through 3 show restricted models. Wald tests for the completeness of the restricted models against the full model indicate that the restricted models in columns 1 and 2 are underspecified.[38] Across the table, I report the estimated (adjusted) variances of teacher quality for each teacher type using each specification. Also, from each restricted model, I estimate the vectors of teacher fixed effects for the relevant teacher types and compare them to their analogs from the complete model in column 4 by reporting correlation coefficients.

Table C.1 shows that as the specifications become progressively richer (moving from left to right in the table), the estimated variances of teacher quality for the different teacher types generally decline.[39] There is a considerable drop in the variance estimates moving from specification (2) to specification (3), where the school-level covariates and fixed effects are included. This drop may reflect a reduction in omitted variables bias, but may also reflect the removal of any across-school variation in teacher quality generated by teacher sorting. Unfortunately, outside of a controlled experiment, there is no clear way to disentangle across-school differences in teacher

[38] 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 variable groups B and C in each specification. 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 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 Table C.1). 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 the table) and merits additional attention in future research.

[39] The one exception is for math teachers moving from column 3 to column 4 in both tested subjects.

quality from the other differences across schools that might affect student performance. By focusing within schools and students, the full model in column (4) ensures that teacher effects will not be biased by school-level factors that influence student achievement. However, estimates from the full model may understate the total variance of teacher quality for each teacher type by omitting any across-school variance.

The correlation coefficients relating the estimated teacher effects across the different models provide one gauge of the extent to which omitted variables bias influences teacher-fixed-effect estimates. The reported correlations indicate that the restricted models can significantly misrepresent teacher rankings.[40]

[40] Harris and Sass (2006) report a similar result. However, in addition to changes in the magnitude of the omitted variables bias moving across specifications, teacher rankings may also be changing because the richer models effectively narrow teachers' comparison groups. If teachers are heavily sorted across schools and across students (within schools), these comparison-group shifts will have larger effects on teacher rankings. Of note, the dispersion analysis in Section III and Appendix B does not find evidence of strong student-teacher sorting, at least within schools.

# Appendix D
# 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 VI 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:

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

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.[41] Given these conditions I can rewrite equation (D.1) as:

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

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

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

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

[41] If these covariances were negative, the procedure outlined in this appendix would not estimate an upper bound and the correlation coefficient could potentially be even higher than is reported here. Because the estimated upper bound on the correlation coefficient is greater than 1, these covariances cannot be negative.

(D.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:

(D.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 IV. 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 E
# Quantitative Properties of the Stanford 9 Exams in Secondary School

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

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.

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 here controls for this 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 test-score ceiling in math on the estimated variances 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 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 the estimated variances of math-teacher quality in columns 3 and 4 of Table C.1 indicates that the former effect dominates. One explanation for this result is that the degree of student-teacher sorting among math teachers is higher for students in lower achievement deciles.[42] This may not be the case for student-teacher sorting in social studies.

[42] 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 math teachers who teach lower-achieving students.

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 ceiling effect does not have a noticeable impact on the estimated variance of English-teacher quality moving from column (3) to column (4) in Table C.1 (such that it breaks away from the downward trend). However, for math teachers, the variance estimate slightly increases once the student fixed effects are added to the model. Again, this may reflect a higher degree of student-teacher sorting among math teachers for students in lower achievement deciles relative to those in higher 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.

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

Hamilton, Barton H., Jack A. Nickerson and Hideo Owan, "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111:3 (2003), pp. 465-497.

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 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 WP 07-08 (2007).

Mas, Alexandre and Enrico Moretti, "Peers at Work," NBER WP 12508 (2006).

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.