An Empirical Analysis of Teacher Spillover Effects in Secondary School

Cory Koedel*
University of Missouri

February 2009

This paper examines whether educational production in secondary school involves joint production among teachers across subjects. In doing so, it also provides insights into the reliability of value-added modeling. Teacher value-added to reading test scores is estimated for four different teacher types: English, math, science and social studies. The initial results indicate that reading output is jointly produced by math and English teachers. However, while falsification tests confirm the English-teacher effects, they cast some doubt about whether the math-teacher effects are free from sorting bias. The results offer a mixed review of the value-added methodology, suggesting that it can be useful but should be implemented cautiously.

*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, Shawn Ni, Jesse Rothstein, seminar participants at UC San Diego, UC Riverside, RAND Corporation, Southern Methodist University, Florida State University, Michigan State University, the University of Missouri and two anonymous referees 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.

The structure of secondary-level education, where students are taught by multiple teachers each year, allows for the possibility of joint production among teachers. This paper uses value-added modeling to evaluate the extent to which teacher quality spills over across subjects in secondary school. Despite annual expenditures in excess of half a trillion dollars on public elementary and secondary education in the United States, this aspect of educational production has been virtually ignored by both researchers and policymakers.[1]

The vast majority of the recent work on teacher quality has focused on elementary-school teachers.[2] This is not surprising – student-teacher assignments in elementary school are one-to-one, are unlikely to be plagued by the degree of sorting bias found in secondary school (where students can be sorted across subjects in addition to teachers), and standardized tests in the one-size-fits-all mold will better measure the performance of younger and less ability-stratified students. All of these factors make for more tractable empirical analyses at the elementary level. However, because the schooling structure in secondary school differs markedly from that in elementary school, elementary-level analyses may to be ill-suited to inform many of the policy relevant questions about educational production in secondary school.

Although researchers and policymakers have generally assumed that student progress in one subject is unrelated to teacher quality in others in secondary school, this assumption is not supported by any empirical evidence. Using a detailed value-added approach to modeling student achievement, I examine which teacher inputs in secondary school affect reading output.[3]

[1] Expenditure estimate from the *Public Education Finances* report issued in April of 2006 by the United States Census Bureau.

[2] See, for example, Harris and Sass (2006), Hanushek, Kain, O'Brien and Rivkin (2005), Koedel and Betts (2007), Nye, Konstantopoulos and Hedges (2004), Rivkin, Hanushek and Kain (2005), and Rockoff (2004). One notable exception is Aaronson, Barrow and Sander (2007)

[3] Because of concerns that a one-size-fits-all standardized math test is inadequate to measure student progress given the subject-specific math curriculums found at the secondary level, I focus only on student reading performance here. Furthermore, the math portion of the Stanford 9 exam for secondary-school students has quantitative properties that are disconcerting whereas the reading exam does not (the Stanford 9 is the exam used for this

I consider the effects of four different teacher types: English, math, science and social studies. The initial analysis indicates that both English and math teachers affect reading test scores, suggesting that educational output in secondary school is jointly produced. However, while post-estimation falsification tests confirm the English-teacher effects, they suggest that the estimated math-teacher effects may contain some sorting bias.

Over the course of evaluating teacher spillover effects in secondary school, this study also provides insights into the reliability of value-added modeling. It extends recent work by Rothstein (2008), who raises concerns about the validity of the value-added approach. Rothstein shows that the teacher effects estimated from value-added models are heavily contaminated by student-teacher sorting bias, negating their validity. The results here are much more upbeat, although mixed. On the one hand, as mentioned above, there is suggestive evidence that the value-added methodology does not entirely mitigate sorting bias for math teachers. However, on the other hand, there is no evidence of sorting bias in the effects estimated for English teachers. These mixed results suggest that the value-added methodology can be useful, but should be implemented cautiously.

The remainder of the paper is organized as follows: Section 1 develops a value-added framework for estimating teacher effects in secondary school, Section 2 describes the data, Section 3 details the methodology for evaluating the teacher effects, Section 4 presents initial results, Section 5 performs falsification tests and Section 6 concludes.

1. Teacher Effects and the Educational Production Function

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

analysis – details on the quantitative properties of the exam are available upon request). Nonetheless, results from a similar analysis based on math test scores are available upon request.

unavailable, researchers have focused on estimating educational production in terms of value-added where a lagged performance measure is used as a substitute for the full history of inputs that affect student performance prior to year *t*.[4] The general value-added framework explains current performance as a function of current inputs while controlling for past performance:

$$Y_{isjt} = f(Y_{ijs(t-1)}, \alpha_i, X_{it}, \delta_{is}, S_{it}, C_{it}, \theta_{i1j}, \theta_{i2j}, ..., \theta_{iKj}) \qquad \textbf{(1)}$$

In (1), Y_{isjt} is a test score for student *i* at school *s* with teacher-set *j* in year *t*, α_i represents observed and unobserved time-invariant student characteristics, X_{it} is a vector of time-varying observable student characteristics, including student *i*'s grade level, δ_{is} represents observed and unobserved time-invariant school characteristics, S_{it} is a vector of observed time-varying school characteristics, C_{it} is a vector of time-varying observable classroom characteristics and θ_{ikj} measures the quality of teacher *k* who teaches student *i* and is part of teacher-set *j*. A specific form of this general value-added model, the gainscore model, is also commonly employed in empirical work.

I estimate teacher value-added to reading test scores in secondary school for four teacher types: English, math, science and social studies.[5] An overview of the California content standards for English, math, science and social-studies classrooms at the secondary level (available from the California State Board of Education) indicates that social-studies teachers are the only non-English teachers who would be expected to have an effect on reading performance through direct instruction based on course content. Because the course-content standards for math and science classrooms make no mention of critical reading skills, any effects that these

[4] The ability of the lagged performance measure to appropriately substitute for the full history of inputs in the achievement function is questionable. In fact, it is the potential failure of the lagged performance measure as a substitute for this information that underlies much of the recent concern about the value-added methodology. For further discussion, see Todd and Wolpin (2003) and Rothstein (2008).

[5] These four teacher types are the most common in San Diego high schools and arguably most relevant for evaluating cognitive performance.

teachers might have on student reading performance are more likely to be the result of effects on student effort and motivation.[6]

I index English teachers from j = 1,…,J; math 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 English teacher, *p*th math teacher, *q*th science teacher and *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 θ_j indicates the quality of English teacher *j*, θ_p indicates the quality of math teacher *p*, and so on. I estimate the effects of these four teacher types on student performance using the following reduced-form value-added specification:

$$Y_{ist}^{jpqr} = \alpha_i + Y_{is(t-1)}^{jpqr}\psi + X_{it}\gamma + D_{it}^{school}\delta_S + S_{it}\rho + C_{it}\eta$$
$$+ D_{it}^{J(English)}\theta_J + D_{it}^{P(math)}\theta_P + D_{it}^{Q(sci)}\theta_Q + D_{it}^{R(soc)}\theta_R + \varepsilon_{it} \qquad \textbf{(2)}$$

All of the vectors of explanatory variables in (2) are defined as above and a list of the sets of controls in each vector is available in Table 1. Indicator variables for schools and teachers are denoted by a "D" and appropriately labeled. Teachers are indexed by subject and denoted by superscripts. Equation (2) allows for teacher spillover effects by allowing multiple teachers to affect student performance.[7]

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 controls for the variety of subject-materials to which students are exposed in high school, and for across-subject student sorting, which is generally not an issue in elementary school but may be very important in secondary school. To control for peer effects, the model

[6] This assertion, although reasonable, is somewhat speculative. Empirically, I do not restrict the mechanisms by which the different teacher-types might affect student performance in secondary school.

[7] I also consider models that include observable teacher characteristics, like experience. The results from these models are virtually identical to those presented in the text. As has been found in other studies, teacher value-added is explained only marginally, at best, by observable teacher qualifications.

includes information on the year-(t-1) achievement of classroom-level peers for each student's math and English classrooms. Also, class-size controls are included to prevent variation in class size from being misinterpreted as variation in teacher quality (again in math and English classrooms).

The specification in (2) minimizes omitted variables bias generated by unobserved heterogeneity in student ability across teachers because teacher effects are estimated relative to other teachers who teach the same students. As long as student-teacher sorting is based on time-invariant factors (e.g., underlying ability, or perhaps parental lobbying), the estimated teacher effects will not be biased by differences in student ability across teachers. However, if students are sorted into classrooms based on time-varying as well as time-invariant characteristics, estimates from (2) may still be biased. This issue will be re-visited in Section 5.

If teacher quality varies across schools or across students within schools, estimates from (2) will understate the total variance of teacher quality in secondary schools because the model ignores any between-school and between-student variance in teacher quality. In an omitted analysis, I consider different forms of the value-added model in (2) that allow for across-student and across-school variation in teacher quality. I find little evidence that the within-school, across-student component of the variance of teacher quality is large. However, across-school variation in teacher quality may be of a non-negligible magnitude.[8] The estimates from equation (2) may understate the variance of teacher quality in secondary school through their omission of this across-school variance.

[8] 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. However, these other models allow for the *possibility* that there is non-negligible across-school variance in teacher quality.

I adopt the method of Anderson and Hsiao (1981) to estimate the model in (2). This method involves first differencing the model 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 $(Y^{jpqr}_{is(t-1)} - Y^{jpqr}_{is(t-2)})$ with $(Y^{jpqr}_{is(t-2)})$. 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 year-(t-1) test score is a direct function of $\varepsilon_{i(t-1)}$. The first-differenced version of equation (2) is detailed below:

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

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.[9] The key assumption required for the instrumentation to be valid is that the error terms in equation (2) are serially uncorrelated (such that the year-(t-2) test score is uncorrelated with the first-differenced error term). Although this assumption is not directly verifiable using equation (2), 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.[10,11]

[9] The year-(t-2) test-score level is a powerful instrument: its t-statistic is above 50 in the first stage.

[10] I evaluate the white noise assumption for the error terms by measuring the serial correlation between the first-differenced error terms, within students, in the first-differenced version of equation (2). The individual ε_{it}'s are serially uncorrelated if the first-differenced error terms are serially correlated with a magnitude of -0.5. For students where more than one first-differenced equation is estimated, the serial correlation between the first-differenced error terms is approximately -0.45. This correlation estimate will be biased toward zero because it is based on estimates of the first-differenced error terms.

[11] I cluster standard errors at the student level per the previous footnote.

One final concern with the estimation of equation (2) is that strict tracking of students to teachers may prevent the multiple teacher effects from being identified. Essentially, what is required for identification is that students are "sufficiently" dispersed across teachers in different subjects; or, put differently, that students are not heavily tracked across subjects. It is straightforward to show that only very mild student dispersion across teachers in secondary school is mechanically required to identify the multiple teacher effects. The next section provides empirical evidence showing that student sorting is relatively mild within San Diego secondary schools. The level of student dispersion across teachers in different subjects is more than sufficient for the identification of the multiple teacher effects.

2. 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. 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.[12]

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.[13]

The modeling structure in equation (2) 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.[14] Science and social-studies classes are not taken each year by most students, making a similar restriction based on these teachers infeasible. Instead, the value-added model includes indicator variables for whether each student took a science or social-studies class in each year. As is the case with English and math, the model also includes

[12] Eighth-grade test-scores are used only as year-(t-2) explanatory variables in the final models.
[13] 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.
[14] 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. Many of the students who are not assigned to an English teacher are designated as English learners. Also note that a small fraction of students took more than one math or English class in any given year. I included these students in the analysis and assigned equal weight to both teachers in that subject-year. For example, if a student took two full English classes in a given year, each teacher would be assigned half credit for that student. In the data, this scenario would be indistinguishable from a student who split a single English class between two teachers, in which case each teacher would also be assigned half credit.

controls for which types of social-studies and science classes are taken by students. By grade level, Table 2 details the class-taking behavior of the final student sample. Appendix A provides summary statistics showing that the sample is slightly advantaged relative to the entire student population at SDUSD but generally representative.

For teachers, I expect sampling variation to have a significant impact on the estimated teacher effects by analogy to Kane and Staiger's analysis of school quality (2002). Thus, although I include indicator variables in the model for all teacher assignments, I only analyze teachers who teach at least 20 students from the restricted student sample.[15] The results presented below are not sensitive to reasonable adjustments to this threshold. The final dataset includes over 1000 teachers who teach at least 20 students in their respective subjects and more than 58,000 test-score records from over 17,000 different students.[16] Because the final samples of students and teachers are likely to be more homogeneous than their respective populations given the inclusion restrictions, the results may understate the variance of teacher quality in secondary school.

Finally, I evaluate student-teacher sorting, or ability grouping, at SDUSD. First, I calculate the average within-teacher standard deviation of students' year-(t-1) test scores and compare it to analogous measures based on simulated student-teacher matches that are either randomly generated or perfectly sorted. If the average actual within-teacher standard deviation differs from the average within-teacher standard deviation estimated from the simulated random

[15] I do not include indicator variables for teachers to whom less than five students are assigned. There are very few such teachers in any subject. Also, by only analyzing teachers who teach more than 20 students, this creates a small non-overlapping set of students in the math and English classrooms of the teachers whose effects are part of the analysis below. For example, a student might take an English class with a 20-plus teacher and a math class with a teacher who teaches less than 20 students. Although the model estimates effects for both teachers, the English-teacher effect in this case would included in the teacher-effect analysis, but the math-teacher effect would not be. I minimize this non-overlapping set in the data by initially focusing only on students who are assigned to both a math and English teacher in each year of the data panel (see discussion above).

[16] By subject, there are 381 English teachers, 290 math teachers, 221 science teachers and 197 social-studies teachers who teach at least 20 students in the data panel.

assignment, it would suggest some degree of ability grouping. This approach follows Aaronson, Barrow and Sander (2007). Table 3 shows the results for each teacher type. The estimates are presented as ratios of the standard deviation of interest to the average within-grade standard deviation of the test (weighted across grades, calculated using my sample). The table shows that although students do not appear to be randomly assigned to teachers, the assignment pattern is much closer to what would be expected from random assignment than from perfect sorting. This implies that students are not strictly tracked, at least based on test scores, at SDUSD.

I also use teacher-by-teacher Herfindahl indices to evaluate student tracking across subjects. The Herfindahl indices provide a more general measure of student dispersion across teachers than the results in Table 3, which are based entirely on test-score performance. I calculate Herfindahl indices for each teacher type going into the classrooms of each other teacher type and use group-level averages to measure student dispersion. For example, the Herfindahl index for math teacher *j* sending students into the classrooms of social studies teachers r = 1,…,R can be written as $H = \sum_{r=1}^{R} (S_{rj} / S_j)^2$ where S_{rj} is the share of math-teacher *j*'s students who are taught by social-studies-teacher *r* and S_j is the total number of students taught by math-teacher *j*. The average index values from the twelve sets of indices (i.e., math-to-English, math-to-science,…, English-to-math, etc.) range from roughly 0.10 to 0.20. To put these numbers in context, an index value of 0.15 would suggest that the average sending teacher could send, at most, approximately 35 percent of her students to a single receiving teacher. Alternatively, she could send 15 percent of her students to six different receiving teachers, and the remaining 10

percent to a seventh teacher. Overall, the Herfindahl indices corroborate the evidence from Table 3 by showing that students are not strictly tracked across classrooms at SDUSD.[17]

3. Methods

I describe the variance of the distribution of teacher quality for each teacher type to convey the importance of differences in teacher quality, by subject, in determining reading performance in secondary school. First, I perform Wald tests for the null hypothesis that the variation in teacher quality for each teacher type is equal to zero:

$$H_0 : \theta_1 = \theta_2 = ... = \theta_J = \bar{\theta}$$
$$W = (\hat{\theta} - \hat{\bar{\theta}}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \hat{\bar{\theta}}\ell_J) \qquad \textbf{(3)}$$

In (3), $\hat{\theta}$ is the Jx1 vector of estimated teacher fixed effects, $\hat{\bar{\theta}}$ is the sample average of the $\hat{\theta}_j$'s, $\hat{V}_J$ is the JxJ portion of the estimated variance matrix corresponding to the teacher effects being tested and ℓ_J is a Jx1 vector of ones.[18] Under the null hypothesis, W is distributed $\chi^2_{(J-1)}$.

Although the Wald tests are useful for determining statistical significance, they do not provide an indication of *economic significance*. Therefore, I also empirically estimate the variance of teacher quality for each teacher type. To do this, I first calculate the total fixed-effects variance. For English teachers, this variance is:

$$Var(\hat{\theta}) = \left(\frac{1}{J-1}\right)\sum_{j=1}^{J}[\hat{\theta}_j^{(English)} - (1/J)\sum_{j=1}^{J}(\hat{\theta}_j^{(English)})]^2 \qquad \textbf{(4)}$$

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 (4) overstates

[17] I calculate the group-level averages based on a random draw of 50 sending teachers in each subject. All of the Herfindahl-index calculations are available upon request.

[18] 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 little effect on my results.

the variance of teacher quality because it includes the variance of the estimation error. I define the estimation-error variance as $Var(\lambda)$ and the variance of the teacher-quality signal, the outcome of interest, as $Var(\theta)$. To separate the estimation-error variance from the variance of the teacher-quality signal, I first assume that $Cov(\theta, \lambda) = 0$.[19] This allows for the total variance of the teacher fixed effects to be decomposed as follows:

$$Var(\hat{\theta}) = Var(\theta) + Var(\lambda) \quad \textbf{(5)}$$

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:

$$(\frac{1}{J-1}) * [(\hat{\theta} - \hat{\bar{\theta}}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \hat{\bar{\theta}}\ell_J)] \approx \frac{Var(\hat{\theta})}{Var(\lambda)} \quad \textbf{(6)}$$

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

$$(\hat{\theta} - \hat{\bar{\theta}}\ell_J)'(\hat{V}_J)^{-1}(\hat{\theta} - \hat{\bar{\theta}}\ell_J) = \frac{(\hat{\theta}_1 - \hat{\bar{\theta}})^2}{\hat{\sigma}_1^2} + \frac{(\hat{\theta}_2 - \hat{\bar{\theta}})^2}{\hat{\sigma}_2^2} + ... + \frac{(\hat{\theta}_J - \hat{\bar{\theta}})^2}{\hat{\sigma}_J^2} \quad \textbf{(7)}$$

In (7), $\hat{\sigma}_j^2$ is the square of the standard error estimate for the effect of teacher *j*. Thus, scaling the Wald statistic 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 (5) and (6). 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:

$$Var(\theta) = Var(\hat{\theta}) - (Var(\hat{\theta}) / A) \quad \textbf{(8)}$$

I use estimates from equation (8) to evaluate the effects of distributional shifts in teacher quality on student performance for each teacher type.

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

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 their approach if instead of using equation (6), I estimated the ratio of the total-fixed-effects variance to the estimation-error variance as:

$$\frac{Var(\hat{\theta})}{Var(\lambda)} \approx \frac{(1/J)*\sum_{j=1}^{J}(\hat{\theta}_j - \bar{\theta})^2}{(1/J)*\sum_{j=1}^{J}(\hat{\sigma}_j^2)} \qquad \textbf{(9)}$$

Although equation (9) 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 (9) is no longer directly tied to the Wald statistic. The appeal of my approach is that the variance estimates are directly tied to the Wald statistic through equation (6).

4. Initial Results

Based on estimates from the student-achievement specification in (2), and assuming the model's validity, I evaluate the effects of variation in teacher quality on student reading performance in secondary school. First, I perform Wald tests of the form in equation (3) for each teacher type. Table 4 documents the results from these tests. The first row of the table shows estimates from a basic model that includes only English teachers and the subsequent rows consider the inclusion of all possible teacher combinations.

Table 4 suggests that variation in teacher quality among English, math and social-studies teachers all affect student performance in reading. However, the table ignores the possibility of interaction effects among the different teacher types. I explore this possibility by adding interactions between English and math teachers, English and social-studies teachers, and math

and social-studies teachers to model (5) from Table 4. To maintain consistency with the preceding analysis, I only analyze teacher interactions that affect at least 20 students.[20] Interactions between English and social-studies teachers and math and social-studies teachers are jointly insignificant in the model. However, interactions between English and math teachers are significant at the one-percent level of confidence. Furthermore, the inclusion of the English- and math-teacher interactions results in the social-studies-teacher indicators becoming statistically insignificant.[21] This result is quite robust and 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 for social-studies teachers to fall to the point of statistical insignificance).[22] This suggests that given a typical six-class schedule, students' social-studies teachers are strong predictors of their math and English teacher combinations, but it is these teacher combinations that are predicting student performance.

Ultimately, the Wald tests show that the appropriate reading-achievement specification includes indicator variables for just English and math teachers as well as indicators for the interactions between these two teacher types (there are 404 interactions that affect 20 or more students). It excludes both science and social-studies teachers. Table 5 details this final reading-achievement specification, which suggests that math-teacher effects spill over across subjects.

[20] Unlike for the individual-teacher indicators, I do not include interaction indicators in the model if there are less than 20 shared students. This means that the interaction effects are estimated relative to a catch-all group of students whose math and English teachers share less than 20 students. One avenue through which an interaction effect would be expected would be teacher teamwork across subjects, which would be less likely to occur with fewer shared students.

[21] The p-value on this new Wald statistic for the inclusion of the social-studies teacher indicator variables exceeds 0.90.

[22] In addition to the possibility that the interaction effects are causal, it is also possible that they serve only as a proxy for student-teacher sorting that is not otherwise captured by the model. The next section more thoroughly explores the possibility of sorting bias in the model, although I am unable to verify the specific source of the interaction effects.

Table 6 evaluates the economic significance of the teacher effects given the statistical significance results from Table 5. It presents estimates from the basic model, which includes only English teachers, as well as the full model that incorporates the math-teacher effects and the interaction effects. The table reports the unadjusted raw variance of the teacher fixed effects and the adjusted variance for each teacher type and for the interaction effects. The variance estimates are presented as ratios of the standard deviation of the teacher quality distribution to the weighted average of the within-grade standard deviations of test scores (calculated using my sample, where the weights correspond to the sample size in each grade). For example, in the basic model in the first vertical panel of Table 6, a one-standard-deviation increase in English-teacher quality (adjusted) corresponds to a 0.10 average within-grade standard deviation improvement in test scores.

The estimates from Table 6 indicate that differences in teacher quality in secondary school have non-negligible effects on student performance. Furthermore, math-teachers' spillover effects vary nearly as much as the direct effects of English teachers – the estimated effect of a one-standard-deviation improvement in math-teacher quality corresponds to a 0.06 standard-deviation improvement in reading test scores compared to the 0.07 standard-deviation improvement that would accompany a similar move in the English-teacher-quality distribution.

5. Falsification Tests

The analysis thus far has closely followed the recent value-added literature. In fact, with the controls for student fixed effects and the subject-level indicator variables, the model in (2) is among the most detailed value-added specifications available. However, recent research on the validity of value-added modeling by Rothstein (2008) suggests that the estimates presented in the previous section may still be biased by student-teacher sorting. Rothstein identifies two types of

sorting relevant to the identification of unbiased teacher effects – static and dynamic. Static sorting refers to sorting based on time-invariant student characteristics while dynamic sorting refers to sorting based on time-varying characteristics. Many of these characteristics are likely to be unobserved by the econometrician. Rothstein shows that while econometric techniques are sufficient to estimate unbiased teacher effects given static sorting, they will generally be insufficient when students are sorted to teachers dynamically. In the analysis here, note that equation (2) controls for static sorting by including student fixed effects. However, mechanically, the first-differencing of the equation requires that students' current teacher assignments are uncorrelated with both the current and lagged disturbance terms, ε_{it} and $\varepsilon_{i(t-1)}$. This would be clearly violated if students were sorted to teachers based on time-varying characteristics.

Ex ante, it is unclear whether we should expect a correlation between students' current-teacher assignments and the error terms in the first-differenced version of equation (2). Expectations about whether such a correlation exists depend on assumptions about how students are assigned to teachers. Without strong priors on the student-teacher assignment process, validating the value-added model is entirely an empirical exercise.

Based on Rothstein (2008), I perform falsification tests of the results from the previous section. The falsification tests are straightforward - because future teachers cannot possibly have causal effects on current student performance, I examine whether the value-added model estimates non-zero "effects" for future teachers. Any such effects can only reflect sorting bias, suggesting that the value-added model has inadequate controls and that all of the teacher effects from the model are potentially biased. To begin, I replicate Rothstein's analysis as closely as possible using my dataset. I start with a sample of tenth-grade students in San Diego (a

subsample of the student population used to generate the results in the previous section) and estimate the following model of student achievement:

$$\Delta Y_i^{10} = S_i\delta + T_i^{9e}\gamma^{9e} + T_i^{10e}\gamma^{10e} + T_i^{11e}\gamma^{11e} + T_i^{9m}\gamma^{9m} + T_i^{10m}\gamma^{10m} + T_i^{11m}\gamma^{11m} + \varepsilon_i \qquad \textbf{(10)}$$

Equation (10) is a gainscore model and corresponds to Rothstein's equation (17). ΔY_i^{10} represents a student's test-score gain going from the ninth to tenth grades, S_i is a vector of school indicator variables, and T_i^{xy} is a vector of teacher indicator variables for student *i* in grade *x* and subject *y* (where *y* = *e* or *m*, for English or math). Correspondingly, γ^{xy} is a vector of teacher effects corresponding to the set of teachers who teach in grade *x* and subject y. Rothstein's basic argument is that if the vectors of future teacher effects are non-zero (for eleventh-grade teachers in this case), the teacher effects from the model are not exogenous and therefore cannot be given a causal interpretation.

I estimate equation (10) using a sample of students selected to match Rothstein's sample as closely as possible to allow for a straightforward comparison. Rothstein is very specific in his sample design for numerous reasons that he outlines in his paper. Of particular relevance, Rothstein focuses on a single cohort of students passing through the North Carolina public schools who do not switch schools and who have observed test-score data in three consecutive years, in addition to observed teacher assignments in a fourth consecutive year. Table 7 is comparable to Rothstein's Appendix Table A.2, which details how he arrives at the sample of students that he uses to estimate his version of equation (10). Although there are differences between Table 7 and Rothstein's Table A.2, these differences are largely the result of differences in our data sources and should not affect the comparability of results.

Table 8 details the results from my replication of Rothstein's analysis. The table reports the adjusted and unadjusted variance of the teacher effects for current and future teachers from

equation (10). I follow Rothstein's approach to reporting the teacher-effect variances, borrowed from Aaronson, Barrow and Sander (2007), where the unadjusted variance is just the raw variance of the teacher effects and the adjusted variance is equal to the raw variance minus the average of the square of the robust standard errors. The variance estimates are reported as ratios of the standard deviation of the distribution of teacher effects to the standard deviation of test scores. I follow the instructions provided in Rothstein's Appendix A to estimate the within-school variance of teacher effects without teachers switching schools.

Despite the simplicity of the value-added model in (10), the results in Table 8 provide useful information. First, for English teachers, the adjusted-variance estimates indicate that eleventh-grade English teachers have non-zero "effects" on tenth-grade reading performance, suggesting a non-negligible sorting bias component to the value-added estimates. However, tenth-grade English teacher effects are significantly more variable. One interpretation of this result is that the value-added coefficients for tenth-grade English teachers from equation (10) are comprised of two signal components – one that is causal and another that reflects sorting bias.[23] For math teachers, using an analogous argument, the table provides no indication that there are causal teacher effects. In fact, the "effects" are larger for eleventh-grade math teachers than for tenth-grade math teachers. This suggests that the math-teacher indicators in equation (10) are merely capturing sorting bias (note that the finding of larger "effects" for eleventh-grade math teachers is consistent with secondary-school students being more stratified by ability across math subjects, and in all likelihood math teachers, in later grades).

To more thoroughly investigate the model-validity issue, I extend the falsification exercise to my primary model in equation (2). Equation (2) includes numerous controls that are omitted from equation (10) and are likely to mitigate sorting bias in the estimated teacher effects.

[23] These signal components are, of course, in addition to the noise component in the estimated teacher effects.

Furthermore, the single-cohort approach taken by Rothstein (2008), and replicated here in equation (10), is likely to exacerbate "transitory" sorting bias where non-persistent student sorting from year to year biases results.[24] Because the dataset used to estimate the model in (2) spans multiple years for many teachers, it will be helpful to remove any such transitory bias.

My approach to the falsification test in the primary specification is *ad hoc* – I simply add future-teacher indicator variables to the already fully specified model. This approach has two caveats. First, not all of the students in my sample have future teacher assignments as this was not a criterion for inclusion into the original model. Therefore, I only estimate future teacher effects for a subset of the students in the data.[25] Second, adding future teacher effects to the within-students model, which is first-differenced, is complicated because a student's future teacher in the lagged-score model is the same as that student's current teacher in the current-score model. So, for example, a student's tenth-grade math teacher enters into the model for ninth-grade value-added as a future teacher and the model for tenth-grade value-added as a current teacher. A similar problem arises for some of the eleventh-grade teachers. I resolve this issue by allowing year-t teachers to have one "effect" in the year-(t-1) model and a separate effect in the year-t model – that is, I do not difference out the teacher indicator variables. Because the current-score teacher effects may be partially causal, while the lagged-score effects cannot be, this seems most reasonable. The current-score and lagged-score effects are not separately identifiable, but are captured by a single coefficient for each current-year teacher.

[24] As just one example of a source of such bias, a principal may alternate across years in assigning the most troublesome students to the teachers at her school.

[25] Approximately 86 percent of the students in my original sample have future English-teacher assignments and 73 percent have future math-teacher assignments. To maintain consistency with the analysis in the previous section, I only analyze future teacher effects for teachers who teach at least 20 students in the future year (although again, I include future-teacher indicator variables in the model for all future teachers who are assigned to at least five students in the sample). Respectively, 83 and 70 percent of the student sample have future English and math teachers who meet this criterion. The discrepancy is largely explained by the fact that a reasonable portion of the sample of future teachers is from the 12th grade and many students do not take math in the 12th grade as only three years of math are required for high-school graduation.

Equations (11), (12) and (13) show the current-score, lagged-score and differenced versions of the model used in the falsification exercise, respectively.

$$Y_{ist}^{jp} = \alpha_i + Y_{is(t-1)}^{jp}\psi + X_{it}\gamma + D_{it}^{school}\delta_s + S_{it}\rho + C_{it}\eta + D_{it}^{j(English)}\theta_J + D_{it}^{p(math)}\theta_P + [D_{it}^{pj(INT)}]\theta_{PJ} + D_{i(t+1)}^{j(English)}\pi_J + D_{i(t+1)}^{p(math)}\pi_P + \varepsilon_{it} \quad \textbf{(11)}$$

$$Y_{is(t-1)}^{jp} = \alpha_i + Y_{is(t-2)}^{jp}\psi + X_{i(t-1)}\gamma + D_{i(t-1)}^{school}\delta_s + S_{i(t-1)}\rho + C_{i(t-1)}\eta + D_{i(t-1)}^{j(English)}\theta_J + D_{i(t-1)}^{p(math)}\theta_P + [D_{i(t-1)}^{pj(INT)}]\theta_{PJ} + D_{it}^{j(English)}\pi_J + D_{it}^{p(math)}\pi_P + \varepsilon_{it} \quad \textbf{(12)}$$

$$Y_{is(t)}^{jp} - Y_{is(t-1)}^{jp} = \alpha_i + (Y_{is(t-1)}^{jp} - \hat{Y}_{is(t-2)}^{jp})\psi + (X_{it} - X_{i(t-1)})\gamma + (D_{it}^{school} - D_{i(t-1)}^{school})\delta_s + (S_{it} - S_{i(t-1)})\rho + (C_{it} - C_{i(t-1)})\eta + D_{it}^{j(English)}\theta_J + D_{it}^{p(math)}\theta_P + D_{it}^{pj(INT)}\theta_{PJ} + D_{i(t+1)}^{j(English)}\pi_J + D_{i(t+1)}^{p(math)}\pi_P - D_{i(t-1)}^{j(English)}\theta_J - D_{i(t-1)}^{p(math)}\theta_P - D_{i(t-1)}^{pj(INT)}\theta_{PJ} - D_{it}^{j(English)}\pi_J - D_{it}^{p(math)}\pi_P + \varepsilon_{it} \quad \textbf{(13)}$$

The second row of equation (13) shows the teacher indicator variables and corresponding vectors of coefficients from the current-score model (11) and the third row shows the teacher indicator variables and corresponding vectors of coefficients from the lagged-score model (12).[26] The teacher coefficients denoted by θ may contain some causal component, while the coefficients denoted by π cannot possibly contain causal information. Grouping terms, the current-teacher coefficients for English and math teachers j and p, respectively, estimate $(\theta_j - \pi_j)$ and $(\theta_p - \pi_p)$. The coefficients of interest from this specification are the coefficients on the future-teacher indicator variables from year-(t+1) – these coefficients will only reflect sorting bias.

Table 9 details the results from the model in (13) for current and future teachers in math and English. For both teacher types, strict exogeneity cannot be rejected as the vectors of future-teacher indicator variables are jointly insignificant. However, despite this apparent confirmation of the model's validity, the results nonetheless suggest that sorting bias may not be entirely

[26] I omit future interaction effects from equations (11) through (13) because a reasonable fraction of the students do not have both math and English teachers in the future year, limiting inference. Furthermore, this analysis casts some doubt about whether the math-teacher effects are free from sorting bias, which also calls into question the validity of the interaction effects.

mitigated for math teachers. To see this, note that in Table 9 the current English-teacher assignments are statistically significant but the current math-teacher assignments are no longer significant in the model. Interpreting these results is complicated by the non-standard interpretation of the coefficients on the current-teacher assignments in equation (13), but the differences between math and English teachers, in conjunction with the results from Table 8, raise some concern about the extent to which sorting bias might be influencing the math-teacher results.

One explanation for the declining significance of the current math-teacher indicators in Table 9 is that the model in (13) is overparameterized, making it appear as though the current math-teacher assignments are insignificant. High correlations between students' current and future math-teacher assignments would be consistent with this explanation. However, teacher-level Herfindahl indices measuring student dispersion from current to future teachers show that students are more dispersed from current math teachers to future math teachers than they are from current English teachers to future English teachers (the average current-to-future-teachers Herfindahl indices for math and English teachers are 0.13 and 0.18, respectively).[27] Furthermore, the within-subject-across-year Herfindahl indices estimated here are similar in magnitude to the within-year-across-subject indices described in Section 2. Therefore, this explanation for the discrepancy between the English- and math-teacher results in Table 9 seems unlikely.

An alternative explanation is that current math-teacher "effects" partially capture sorting bias, and that the future math-teacher assignments also capture sorting bias, which in turn lowers

[27] This result is partly mechanical, as a higher percentage of students are assigned to a future English teacher – see footnote 25. If a student was not assigned to a future-year teacher in a given subject, she was treated as if she was assigned to a unique bin in the Herfindahl-index calculations. However, also note that math classes in secondary school are not grade-level specific as are English classes. Structurally, this suggests that students will be more dispersed into math classrooms across years.

the sorting-bias contribution to the current-year effects. To the extent that this explanation is correct, the falsification tests cast some doubt about the validity of the math-teacher effects (although they do not refute the possibility that such effects exist). The presence of sorting bias in the math-teacher effects is also consistent with the results from Table 8. Although the estimates for current and future teachers from Table 8 are clearly contaminated by sorting bias for both math and English teachers, the current-teacher variances should also contain the variance of any causal effects. Table 8 provides no evidence of a causal component to the math-teacher effects.

In sum, the falsification tests provide robust evidence of important differences in English-teacher quality in secondary school, and show that unbiased estimates of teacher effects can be uncovered from a thorough value-added model. Given the recent controversy surrounding the reliability of value-added modeling, this is an important finding. For math teachers, the results are less clear. While math teachers may also affect reading performance, the falsification tests suggest that student-teacher sorting bias may not be entirely mitigated by the value-added approach.

6. Concluding Remarks

Students' reading test scores in secondary school are strongly influenced by differences in teacher quality among English teachers. Math teachers may also influence reading performance, but the results are less clear. There is no evidence that teacher quality among science or social studies teachers affects reading achievement. Taken together, these results provide insight for researchers and policymakers interested in incentive design for secondary-school teachers. Although they do not preclude some forms of group-based incentives for

teachers in secondary school (e.g., within subject), the adoption of group-based incentives across subjects is not strongly supported by the empirical evidence.

In addition to evaluating the joint-production question in secondary school, this study also contributes to the growing literature on the validity of value-added modeling. In particular, it extends recent research by Rothstein (2008) who shows that value-added estimates of teacher effects are almost entirely reflective of sorting bias. The results here are more upbeat, although mixed. While there is no evidence of sorting bias in the English-teacher effects, there is suggestive evidence of bias in the math-teacher effects. Overall, although a cautious approach is warranted, the results show that the value-added methodology can be a useful tool for the evaluation of teacher effects.

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 of school, school size, 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, chemistry or physics, etc.)

Table 2. Class-Taking Behavior of the Student Sample by Grade Level

	Ninth Grade	Tenth Grade	Eleventh Grade
Classes Taken			
Math	100%	100%	100%
English	100%	100%	100%
Science	44%	88%	82%
Social Studies	79%	27%	99%
Science and Social Studies	26%	19%	81%

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

Table 3. Average Within-Teacher Standard Deviations of Students' Year-(t-1) Test Scores, by Teacher Type

		Within Schools		Across District	
	Actual	**Random Assignment**	**Perfect Sorting**	**Random Assignment**	**Perfect Sorting**
Teacher-Type					
Math Teachers	0.85	0.96	0.17	0.99	<0.01
English Teachers	0.78	0.95	0.15	0.99	<0.01
Science Teachers	0.85	0.96	0.20	0.99	<0.01
Social Studies Teachers	0.80	0.97	0.21	0.99	<0.01

Note: In the "Perfect Sorting" columns, students are sorted by year-(t-1) test-score levels. 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 4. P-Values from Wald Tests for the Joint Significance of the Teacher Indicator Variables, by Teacher Type

	Statistical Significance for Teacher Indicator Variables by Teacher Type			
Teachers Included	English	Math	Science	Social Studies
1. English Only	**<0.01****	-	-	-
2. English and Mathematics	**<0.01****	**<0.01****	-	-
3. English and Social Studies	**<0.01****	-	-	**<0.01****
4. English and Science	**<0.01****	-	0.34	-
5. English, Mathematics and Social Studies	**<0.01****	**<0.01****	-	**<0.01****
6. English, Mathematics and Science	**<0.01****	**<0.01****	0.36	-
7. English, Social Studies and Science	**<0.01****	-	0.60	**<0.01****
8. English, Mathematics, Social Studies and Science	**<0.01****	**<0.01****	0.37	**<0.01****

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

Table 5. Final Reading Achievement Model and Associated P-Values from Wald Tests

	Statistical Significance for Teacher Indicator Variables by Subject		
Teachers Included	English	Mathematics	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 6. 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	0.16	0.10	0.17	0.07
Math Teachers			0.14	0.06
English-Math Teacher Interactions			0.15	0.06

Table 7. Definition of Data Sample Used to Replicate Rothstein's Analysis

	Number	% of Universe
Universe: All students in the tenth grade in 1999-2000 (with tenth-grade reading test score)	7187	100%
Drop:		
(1) Duplicate observations in any year	-0	0.0%
(2) Missing data (including no student record) in grade 9, 10 or 11	-2274	31.6%
(3) Missing test-score gain in grade 9 (no 8^{th}-grade score)	-709	9.8%
(4) Inconsistent data on race/gender	-13	0.2%
(5) Skipped or held back	-177	2.5%
(6) Changed schools in grades 9, 10 or 11	-306	4.3%
(7) Missing math or English teacher assignment in any grade (or did not take math or English in grades 9, 10 or 11)	-1215	16.9%
(8) School miscoded	-67	0.9%
(9) Small Within-School Sample	-26	0.4%
(10) Less than 2 sample students assigned to teacher in any grade	-26	0.4%
(11) School has only one teacher	-0	0
(12) Teacher dropped in any year due to within-school collinearity	-200	2.8%
Sample	2174	30.2%

Table 8. Standard Deviations of the Distributions of Current and Future Teacher Effects from a Model with Controls for Schools and Past, Current and Future Teachers. Dependent Variable: Tenth-Grade Gainscore

	Wald Statistic (DF)	P-Value	Unadjusted Variance	Adjusted Variance
Grade 10 English Teachers	133 (68)	<0.01	0.31	0.17
Grade 11 English Teachers	118 (66)	<0.01	0.26	0.10
Grade 10 Math Teachers	166 (97)	<0.01	0.32	0.15
Grade 11 Math Teachers	176 (97)	<0.01	0.35	0.19

Table 9. Standard Deviations of the Distributions of Current and Future Teacher Effects from the Model in Equation (2) with additional Controls for Future Math and English Teachers

	Wald Statistic (DF)	P-Value	Unadjusted Variance	Adjusted Variance
Current English Teachers*	406 (299)	<0.01	0.19	0.10
Future English Teachers	158 (191)	0.96	0.09	0.00**
Current Math Teachers*	274 (248)	0.12	0.15	0.04
Future Math Teachers	177 (168)	0.32	0.09	0.02

* Note that these estimates are for the composite effects documented in equation (13).

** Adjusted-variance estimate was negative.

References

Aaronson, D., Barrow L. & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95-135.

Anderson T.W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of American Statistical Association*, 76(375), 598-606.

Betts, J. R., Zau A., & Rice, L. (2003). *Determinants of Student Achievement, New Evidence from San Diego*. Public Policy Institute of California.

Hanushek, E., Kain, J., O'Brien, D. & Rivkin, S. (2005). The market for teacher quality. NBER WP 11154.

Harris, D. & Sass, T. (2006). Value-added models and the measurement of teacher quality. unpublished manuscript, Florida State University.

Ingersoll, G.M., Scamman, J.P., & Eckerling, W.D. (1989). Geographic mobility and student achievement in an urban setting. *Educational Evaluation and Policy Analysis*, 11(2), 143-149.

Kane, T. & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4), 91-114.

Koedel, C. & Betts, J.R. (2007). Re-examining the role of teacher quality in the educational production function. University of Missouri WP 07-08.

Nye, B., Konstantopoulos, S. & Hedges, L.V. (2004) How large are teacher effects?" *Educational Evaluation and Policy Analysis*, 26(3), 237-257.

Rivkin, S., Hanushek, E. & Kain, J. (2005). Teachers, schools and academic achievement. *Econometrica*, 79(2), 417-58.

Rockoff, J. (2004). The impact of individual teachers on student achievement: evidence from panel data," *American Economic Review*, Papers and Proceedings.

Rothstein, J. (2008). Teacher quality in educational production: tracking, decay, and student achievement. unpublished manuscript, Princeton University.

Rumberger, R.W. & Larson, K.A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education*, 107(1), 1 -35.

Todd, P. & Wolpin, K.I. (2003). Toward a unified approach for modeling the production function for cognitive achievement. *Economic Journal*, 113(485), 3-33.

Appendix A
Data Restrictions

The structure of the model in Section 1 requires at least three contiguous test scores per student for full identification, limiting 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. 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 entire SDUSD high school population. 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 Rumberger and Larson, 1998; Ingersoll, Scamman and Eckerling, 1989).

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%	14%
% Asian	22%	29%
% Hispanic	31%	26%
% English Learners	14%	11%
SAT 9 Math Score*	0	0.19
SAT 9 Reading Score*	0	0.19
Avg. Percentage of School on Free Lunch	44%	41%

My final sample includes 17,468 unique students with at least 3 consecutive 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.