DOES STUDENT SORTING INVALIDATE VALUE-ADDED MODELS OF TEACHER EFFECTIVENESS? AN EXTENDED ANALYSIS OF THE ROTHSTEIN CRITIQUE

Cory Koedel
(corresponding author)
Department of Economics
University of Missouri
118 Professional Building
Columbia, MO 65211
koedelc@missouri.edu

Julian R. Betts
Department of Economics
University of California,
San Diego and NBER
9500 Gilman Drive
La Jolla, CA 92093-0508
jbetts@ucsd.edu

Abstract

Value-added modeling continues to gain traction as a tool for measuring teacher performance. However, recent research questions the validity of the value-added approach by showing that it does not mitigate student-teacher sorting bias (its presumed primary benefit). Our study explores this critique in more detail. Although we find that estimated teacher effects from some value-added models are severely biased, we also show that a sufficiently complex value-added model that evaluates teachers over multiple years reduces the sorting bias problem to statistical insignificance. One implication of our findings is that data from the first year or two of classroom teaching for novice teachers may be insufficient to make reliable judgments about quality. Overall, our results suggest that in some cases value-added modeling will continue to provide useful information about the effectiveness of educational inputs.

1. INTRODUCTION

Economic theory states that in an efficient economy workers should be paid their value marginal product. Implementing this rule in the service sector is not simple, as it is often not obvious how to measure the output of a white collar worker. Teachers provide an example of this problem: public school teachers' salaries are determined largely by academic degrees and credentials and years of experience, none of which appears to be strongly related to teaching effectiveness.

Perhaps in recognition that teacher pay is not well aligned with teaching quality, President Obama has recently called for greater use of teacher merit pay as a tool to boost student achievement in America's public schools. But in the United States teacher merit pay is hardly a new idea. It has been used for at least a century, but most programs are short lived or survive either by giving almost all teachers bonuses or by giving trivial bonuses to a small number of teachers. Teachers have traditionally complained that principals cannot explain why they gave a bonus to one teacher but not another (Murnane et al. 1991). Opponents of teacher merit pay would raise the question of whether we can reliably measure teachers' value marginal products such that informed merit pay decisions can be made.

The advent of wide-scale student testing, partly in response to the requirements of the federal No Child Left Behind (NCLB) law, raises the possibility that it is now feasible to measure the effectiveness of individual teachers. Indeed, recently developed panel data sets link students and teachers at the classroom level, allowing researchers to estimate measures of "outcome-based" teacher effectiveness.[1] Because test scores are generally available for each student in each year, they lend themselves comfortably to a value-added approach where the effectiveness of teacher inputs can be measured by student test score growth. The conjuncture of President Obama's recent calls for teacher merit pay and the development of panel data sets that provide information on student achievement growth raise the stakes considerably: can we use student testing to reliably infer teaching quality?

In most schools, students are not randomly assigned to teachers. A presumption in value-added modeling is that by focusing on achievement growth rather than achievement levels, the problem of student-teacher sorting bias is resolved because each student's initial test score level is used as a control in the model. The value-added approach is intuitively appealing, and increasing demand for performance-based measures by which teachers can be held accountable—at the federal, state, and district levels—has only fueled

1. For recent examples see Aaronson, Barrow, and Sander 2007, Hanushek et al. 2005, Harris and Sass 2006, Koedel and Betts 2007, Nye, Konstantopoulos, and Hedges 2004, and Rockoff 2004.

the value-added fire.[2] However, despite the popularity of the value-added approach among both researchers and policy makers, not everyone agrees that it is reliable. Could it not be the case that a given teacher either systematically or occasionally receives students whose gains in test scores are unusually low, for reasons outside the teacher's control? Ability grouping would be one source of persistent differences in the types of students across classrooms. Random variations, accompanied by mean reversion, would be a source of fleeting differences that a value-added model (VAM) might wrongly attribute to a given teacher.

Recent research by Rothstein (2010) shows that future teacher assignments have non-negligible predictive power over current student performance in value-added models, despite the fact that future teachers cannot possibly have causal effects on current student performance. This result suggests that student-teacher sorting bias is not mitigated by the value-added approach. Rothstein's critique of the value-added methodology comes as numerous studies have used and continue to use the technique. It raises serious doubts about the value-added methodology just as other work, such as Kane and Staiger (2008), Jacob and Lefgren (2007), and Harris and Sass (2007), appears to confirm that value added is a meaningful measure of teacher performance.

We further explore the reliability of value-added modeling by extending Rothstein's analysis in two important ways. First, Rothstein estimates teacher effects using only a single year of data for each teacher. We consider the importance of using multiple years of data to identify teacher effects. If the sorting bias uncovered by Rothstein is transitory to some extent, using multiple cohorts of students to evaluate teachers will help mitigate the bias.[3] For example, a principal may alternate across years in assigning the most troublesome students to the teachers at her school, or teachers may connect with their classrooms more in some years than in others.[4] These types of single-year idiosyncrasies will be captured by single-year teacher effects but will be smoothed out if estimates are based on multiple years of data. Second, we evaluate the Rothstein critique using a different data set. Given that the degree of

2. NCLB legislation is one example of this demand at the federal level (e.g., adequate yearly progress), and states such as Florida, Minnesota, and Texas have all introduced performance incentives for teachers that depend to some extent on value added. For a further discussion of the performance-pay landscape, particularly as it relates to teachers, see Podgursky and Springer (2007).
3. Rothstein notes this in his appendix, although he does not explore the practical implications in any of his models.
4. In addition, some of what we observe to be sorting bias may be attributable to the random assignment of students to teachers across small samples (classrooms). In an omitted analysis, we perform a Monte Carlo exercise to test for this possibility. Although any given teacher may benefit (be harmed) in any given year from a random draw of high-performing (low-performing) students, we find no evidence to suggest that this would influence estimates of the distribution of teacher effects.

student-teacher sorting may differ across different educational environments, his results may or may not be replicated in other settings.

Our extension of Rothstein's analysis corroborates his primary finding—VAMs of student achievement that focus on single-year teacher effects will generally produce biased estimates of value added. However, when we estimate a detailed VAM and restrict our analysis to teachers who teach multiple classrooms of students, we find no evidence of sorting bias in the estimated teacher effects. Although this result depends on the degree of student-teacher sorting in our data, it suggests that at least in our setting, sorting bias can be almost completely mitigated using the value-added approach and looking across multiple years of classrooms for teachers.

Our results in this regard are encouraging; however, less detailed VAMs that include teacher effect estimates based on single classroom observations fare poorly in our analysis. That some VAMs will be reliable but not others, and that value-added modeling may be reliable only in some settings, are important limitations. They suggest that in contexts such as statewide teacher accountability systems, large-scale value-added modeling may not be a viable solution. Because the success of the value-added approach will depend largely on data availability and the underlying degree of student-teacher sorting in the data (much of which may be unobserved), post-estimation falsification tests along the lines of those proposed by Rothstein (2010) will be useful in evaluating the reliability of value-added modeling in different contexts.

Although we do not uncover a well-defined set of conditions under which value-added modeling will universally return causal teacher effects across different schooling environments (outside of random student-teacher assignments, such conditions are unlikely to exist), we do identify conditions under which value-added estimation will perform *better*. The most important insight is that teacher evaluations that span multiple years will produce more reliable measures of teacher effectiveness than those based on single-year classroom observations. Often implicitly, the value-added discussion in research and policy revolves around single-year estimates of teacher effects. Our analysis strongly discourages such an approach.

The remainder of the article is organized as follows. Section 2 briefly discusses the Rothstein critique. Section 3 describes our data set from the San Diego Unified School District (SDUSD). Section 4 replicates a portion of Rothstein's analysis using the San Diego data. Section 5 details our extended analysis of value-added modeling and presents our results. Section 6 shows that transitory sorting bias appears to be driving our findings. Section 7 uses our results to estimate the variance of teacher effectiveness in San Diego. Section 8 concludes.

2. THE ROTHSTEIN CRITIQUE

Rothstein raises concerns about assigning a causal interpretation to value-added estimates of teacher effects. His primary argument is that teacher effects estimated from VAMs are biased by nonrandom student-teacher assignments and that this bias is not removed by the general value-added approach or by standard panel data techniques. Consider a simple VAM of the general form:

$$Y_{it} = Y_{it-1}\delta + X_{it}\beta + T_{it}\theta + (\alpha_i + \varepsilon_{it}). \quad (1)$$

In equation 1, Y_{it} is a test score for student i in year t, X_{it} is a vector of time-varying student and school characteristics (for the school attended by student i in year t), and T_{it} is a vector of indicator variables indicating which teacher(s) taught student i in year t. This model could be reformulated as a gain score model by forcing the coefficient on the lagged test score to unity and moving it to the left-hand side of the equation. The error term is written as the sum of two components, one that is time invariant (α_i) and another that varies over time (ε_{it}).

Rothstein discusses sorting bias as coming from two different sources in this basic model. First, students could be assigned to teachers based on "static" student characteristics. This type of sorting corresponds to the typical tracking story—some students are of higher ability than others, and these students are systematically assigned to the best teachers. Static tracking may operationalize in a variety of ways, including administrator preferences, parental preferences, or teacher preferences (assuming that primary-school-aged children, upon whom we focus here, are not yet able to form their own preferences). Given panel data, the typical solution to the static tracking problem is the inclusion of student fixed effects, which control for the time-invariant components to the error terms in equation 1. If student-teacher sorting is based only on static student characteristics, this approach will be sufficient.

However, the student fixed effects solution to the static tracking problem necessarily imposes a strict exogeneity assumption. That is, to uncover causal teacher effects from a model that controls for time-invariant student characteristics, it must be the case that teacher assignments in all periods are uncorrelated with the time-varying error components in all periods. To see this, note that we could estimate equation 1 by first-differencing to remove the time-invariant component to the error term:[5]

$$Y_{it} - Y_{it-1} = (Y_{it-1} - Y_{it-2})\delta + (X_{it} - X_{it-1})\beta + (T_{it} - T_{it-1})\theta + (\varepsilon_{it} - \varepsilon_{it-1}). \quad (2)$$

5. In the case of first-differencing, it is more accurate to describe the assumption as "local" strict exogeneity in the sense that the error terms across time must be uncorrelated with teacher assignments only in contiguous years.

The first-differencing induces a mechanical correlation between the lagged test score gain and the first-differenced error term in equation 2. This correlation can be resolved by instrumenting for the lagged test score gain with the second-lagged gain, or second-lagged level (following Anderson and Hsiao 1981). In addition, year t teacher assignments may also be correlated with the first-differenced error term. Specifically, if students are sorted *dynamically* based on time-varying deviations (or shocks) to their test score growth trajectories, lagged shocks to test score growth, captured by ε_{it-1}, will be correlated with year t teacher assignments, and the teacher effects from equation 2 cannot be given a causal interpretation.[6] Rothstein's critique can be summarized as follows: If students are assigned to teachers based entirely on time-invariant factors, unbiased teacher effects can in principle be obtained from a well-constructed VAM. However, if sorting is based on dynamic factors that are unobserved by the econometrician, value-added estimates of teacher effects cannot be given a causal interpretation.

Rothstein proposes a falsification test to determine whether VAMs produce biased estimates of teacher effectiveness. He suggests simply adding future teacher assignments to the model and testing whether these teacher assignments have nonzero "effects." Future teachers do not causally influence current test scores, which means that any observed effects must be the result of a correlation between teacher assignments and the error terms. Alternatively, if the coefficients on the future teacher indicator variables are jointly insignificant, sorting bias is unlikely to be a major concern for any teacher effects in the model (as this finding would suggest that the controls in the model are capturing the sorting bias that would otherwise confound the teacher effects). In Rothstein's analysis (2010), his most provocative finding is that future teacher assignments have significant predictive power over current student performance. This result suggests that value-added estimates of teacher effects are contaminated by substantial sorting bias.

3. DATA

We use administrative data from four cohorts of fourth-grade students in San Diego (at SDUSD) who started the fourth grade in the school years between 1998–99 and 2001–2. The standardized test that we use to measure student achievement, and therefore teacher value added, is the Stanford 9 mathematics test. The Stanford 9 is designed to be vertically scaled such that a one-point

6. Serial correlation in the epsilons would imply that if year t teacher assignments are correlated with ε_{it-1}, they will also be correlated with ε_{it}, invalidating any VAM even in the absence of static tracking.

gain in student performance at any stage in the schooling process is meant to correspond to the same amount of learning. [7]

Students who have fourth-grade test scores and lagged test scores are included in our baseline data set. We estimate a VAM that assumes a common intercept across students and a second model that incorporates student fixed effects. In this latter model we also require students to have second-lagged test scores. For each of our primary models, we estimate value added for teachers who teach at least twenty students across the data panel and restrict our student sample to the set of fourth-grade students taught by these teachers.[8] In the baseline data set, we evaluate test score records for 30,354 students taught by 595 fourth-grade teachers. Our sample size falls to 15,592 students taught by 389 teachers in the student fixed effects data set. The large reduction in sample size is the result of (1) the requirement of three contiguous test score records per student instead of just two, which in addition to removing more transient students also removes one year-cohort of students because we do not have test score data prior to 1997–98 (that is, students in the fourth grade in 1998–99 can have lagged scores but not second-lagged scores) and (2) requiring the remaining students be assigned to one of the 389 fourth-grade teachers who teach at least twenty students with three test score records or more.[9] We include students who repeat the fourth grade because it is unlikely that grade repeaters would be excluded from teacher evaluations in practice. In our original sample of 30,354 students with current and lagged test score records, only 199 are grade repeaters.

4. REPLICATION OF ROTHSTEIN'S ANALYSIS

Based on details provided by Rothstein in his 2010 article and corresponding data appendix, we first replicate a portion of his analysis using data from the 1999–2000 fourth-grade cohort in San Diego. This replication is meant to establish the extent to which Rothstein's underlying findings are relevant in San Diego.[10] We estimate the following basic VAM:

$$\Delta Y_i^4 = S_i\delta + T_i^3\pi^3 + T_i^4\pi^4 + T_i^5\pi^5 + \varepsilon_i. \tag{3}$$

7. For detailed information about the quantitative properties of the Stanford 9 exam, see Koedel and Betts (2010).
8. This restriction is imposed because of concerns about sampling variation (see Kane and Staiger 2002). Our results are not sensitive to reasonable adjustments to the twenty-student threshold.
9. Only students who repeated the fourth grade in the latter two years of our panel could possibly have had more than three test score records. There are thirty-two students with four test score records in our data set.
10. The replication data sample is roughly a subsample of the student fixed effects data set, but we use different teachers because Rothstein does not require teachers to teach twenty students for inclusion into the model.

Table 1. Standard Deviations of Teacher Effects from a Model with Controls for Past, Current, and Future Teachers. Dependent Variable: Fourth-Grade Gain in Test Score

	Wald Statistic (DF)	*P*-Value	Unadjusted Standard Deviation	Adjusted Standard Deviation
Grade 4 teachers	952 (292)	<0.01	0.40	0.24
Grade 5 teachers	610 (253)	<0.01	0.30	0.15

Notes: The Wald statistics and p-values are from tests for whether all teachers in the given grade have identical effects on student gains in grade 4. The standard deviations refer to the standard deviations of estimated teacher effects, both raw and adjusted, as explained in the text.

Equation 3 is a gain score model and corresponds to Rothstein's VAM1 model with indicators for past, current, and future teacher assignments. ΔY_i^4 represents a student's test score gain going from the third to fourth grade, S_i is a vector of school indicator variables, and T_i^x is a vector of teacher indicator variables for student i in grade x. Correspondingly, π^x is a vector of teacher effects corresponding to the set of teachers who teach grade x. Rothstein's basic argument is that if future teacher effects—for fifth-grade teachers in this case—are shown to be nonzero, then none of the teacher effects in the model can be given a causal interpretation.[11]

We replicate the data conditions in Rothstein (2010) as closely as possible when estimating this model. There are two conditions that seemed particularly important. First, in specifications that include teacher identifiers across multiple grades, Rothstein excludes students who changed schools across those grades in the data. Second, he also focuses on only a single cohort of students passing through the North Carolina public schools. Similar to Rothstein, the data set used to estimate equation 3 does not include any school switchers and is estimated using just a single cohort of fourth-grade students in San Diego.

In our replication we focus on the effects of fourth- and fifth-grade teachers in equation 3. In accordance with the literature that measures the importance of teacher value added, we report the adjusted and unadjusted variance of the teacher effects. We 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. We follow the steps outlined in Rothstein's appendix to estimate the within-school variance of teacher effects without teachers switching schools. Our results are detailed in table 1.

11. Rothstein's analyses (2009, 2010) are quite thorough, and we refer the interested reader to his paper for more details.

The first two columns of table 1 report the results of separate Wald tests of the hypotheses that all grade 4 teachers have identical effects and that all grade 5 teachers have identical effects. Confirming Rothstein's findings, the null hypothesis that grade 5 teachers have an equal effect on students' gains in grade 4 is rejected with a *p*-value below 0.01.

The next two columns show the raw standard deviations of teacher effects and the standard deviations after adjusting for sampling variance. (These are scaled by dividing by the standard deviation of student test scores.) The adjusted standard deviations are 0.24 and 0.15 for grade 4 and grade 5 teachers, respectively. Rothstein's (2010) estimates that are most analogous to those in our table 1, in terms of both the model and the data, are reported in his table 5 (column 2) for the unrestricted model. There he shows adjusted standard deviations of the distributions of grade 4 and grade 5 teacher effects of 0.193 and 0.099 standard deviations of the test, respectively. These results are also replicated virtually identically in his table 2 (column 7), where he uses a larger student sample and excludes lagged year teacher identifiers from the model. Our estimates, which show a larger overall variance of teacher effects, are consistent with past work using San Diego data (Koedel and Betts 2007). The relevant result to compare with Rothstein (2010) is our estimate of the ratio of the standard deviations of the distributions of future and current teacher effects. Rothstein finds that the standard deviation of the distribution of future teacher effects is approximately 51 percent of the size of that of current teacher effects (i.e., 0.099/0.193), whereas in our analysis this number is slightly higher, at roughly 63 percent (0.15/0.24). Our results here confirm Rothstein's finding that future teachers explain a sizable portion of current grade achievement gains and establish that his primary result is not unique to North Carolina.

The results in table 1, and the corresponding results detailed by Rothstein, suggest that student-teacher sorting bias is a significant complication to value-added modeling. Information about the degree of student-teacher sorting in our data will be useful for generalizing our results to other settings. We document *observable* student-teacher sorting in our data by comparing the average realized within-teacher standard deviation of students' lagged test scores to analogous measures based on simulated student-teacher matches that are either randomly generated or perfectly sorted. This approach follows Aaronson, Barrow, and Sander (2007). Although sorting may occur along many dimensions, the extent of sorting based on lagged test scores is likely to provide some indication of sorting more generally. Table 2 details our results, which are presented as ratios of the standard deviation of interest to the within-grade standard deviation of the test (calculated based on our student sample). Note that while there does appear to be some student sorting

Table 2. Average Within-Teacher Standard Deviations of Students' Period (t − 1) Test Scores

		Within Schools		Across District	
	Actual	Random Assignment	Perfect Sorting	Random Assignment	Perfect Sorting
Standard deviations of lagged scores	0.81	0.90	0.32	0.99	<0.01

Notes: 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 randomly generated numbers 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.

based on lagged test score performance in our data set, this sorting is relatively mild.

5. EXTENSIONS TO MORE COMPLEX VALUE-ADDED MODELS

We extend the analysis by evaluating the effects of future teachers using three models that are more commonly used in the value-added literature. These models include a richer set of control measures. We use a general value-added specification where current test scores are regressed on lagged test scores, but note that it is also somewhat common in the literature to use the gain score model (which is used primarily by Rothstein), where the coefficient on the lagged test score is forced to one and the lagged score term is moved to the left side of the equation.

The first model that we consider, and the simplest, is a basic VAM that allows for the comparison of teacher effects across schools:

$$Y_{it} = \psi_t + Y_{i(t-1)}\phi_1 + X_{it}\phi_2 + T_{it}\theta + \varepsilon_{it}. \quad (4)$$

In equation 4, Y_{it} is the test score for student i in year t, ψ_t is a year-specific intercept, X_{it} is a vector of time-invariant and time-varying student-specific characteristics (see table 3), and T_{it} is a vector of teacher indicator variables where the entry for the teacher who teaches student i in year t is set to one. The coefficients of interest are in the vector of teacher effects, θ.

We refer to equation 4 as the basic model. The most obvious omission from the model is school-level information, whether in the form of school fixed effects or time-varying controls. Researchers have generally incorporated this information because of concerns that students and teachers are sorting into schools nonrandomly. This sorting, along with the direct effects of school-level inputs on student achievement (peers, for example), will generate omitted variables bias in the value-added estimates of teacher effects in equation 4.

Table 3. Controls from Value-Added Models

Student-Level Controls (X_{it})	**School-Level (and Classroom-Level) Controls (S_{it})**
English-learner (EL) status	School fixed effects
Change from EL to English proficient	Classroom-level peer performance in year (t − 1)
Expected and unexpected school changer	Class size
Parental education	
Race	Percentage of student body by:
Gender	Race
Designated as advanced student	English learner status
Percentage of school year absent[a]	Free/reduced price lunch status
	School-changer status

[a]The share of days missed by students is sometimes considered endogenous. Fourth-grade students, however, are not likely to have much influence over their attendance decisions.

This leads to the second model that we consider, the within-schools model, which is more commonly estimated in the literature and includes school-level covariates and school fixed effects.

$$Y_{it} = \xi_t + Y_{i(t-1)}\beta_1 + X_{it}\beta_2 + S_{it}\beta_3 + T_{it}\gamma + \nu_{it}. \quad (5)$$

In equation 5, S_{it} includes school indicator variables and time-varying school-level information for the school attended by student i in year t. The controls in the vector S_{it} are detailed in table 3. The benefit of including school-level information is a reduction in omitted variables bias, including sorting bias generated by students and teachers selecting into specific schools. However, the cost of moving from equation 4 to equation 5 is that it is no longer straightforward to compare teachers across schools.[12]

Finally, in our third specification we incorporate student fixed effects. This approach is suggested by Harris and Sass (2006), Koedel (2009), and Koedel and Betts (2007), among many others:

$$Y_{it} = \alpha_i + \lambda_t + Y_{i(t-1)}\tau_1 + X_{it}\tau_2 + S_{it}\tau_3 + T_{it}\delta + u_{it}. \quad (6)$$

The inclusion of the student fixed effects, α_i, allows us to drop from the vector X_{it} time-invariant student characteristics, leaving only time-varying student characteristics. The benefit of the within-students approach is that teacher effects will not be biased by within-school student sorting across teachers

12. Although teacher effectiveness cannot be compared across schools straightforwardly using value-added estimates from equation 5, this may be acceptable from a policy perspective. For example, policy makers may wish to identify the best and worst teachers on a school-by-school basis regardless of any teacher sorting across schools.

based on time-invariant student characteristics (such as ability, parental involvement, etc.). However, as noted in section 2, there are trade-offs. Recall that the student fixed effects model necessarily imposes some form of the strict exogeneity assumption. Equation 6 also narrows teachers' comparison groups to those with whom they share students, meaning that identification comes from comparing test score gains for individual students when they were in the third and fourth grades. In addition, the incorporation of the student fixed effects makes the model considerably noisier.[13] Finally, the size of the student sample that can be used is restricted in equation 6 because a student record must contain at least three contiguous test scores, instead of just two, to be included in the analysis (as described in section 3).

Despite these concerns, econometric theory suggests that the inclusion of student fixed effects will be an effective way to remove within-school sorting bias in teacher effects as long as students and teachers are sorted based on time-invariant characteristics. We estimate the within-students model by first-differencing equation 6 and instrumenting for students' lagged test score gains with their second-lagged test score levels.[14] This general approach was developed by Anderson and Hsiao (1981) and has recently been used by Harris and Sass (2006), Koedel (2009), and Koedel and Betts (2007) to estimate teacher value added.[15] Note that to completely first-difference equation 6 we must incorporate students' lagged teacher assignments, which will appear in the period $(t - 1)$ version of equation 6.[16] That is, the model compares the effectiveness of students' current and previous year teachers.

We start with our baseline student samples—the 30,354 student/595 teacher sample for our basic and within-schools models and the 15,592 student/389 teacher sample for our within-students model. For the students in these samples, we then move forward one year and identify fifth-grade teacher assignments. In each sample, approximately 85 percent of the students appear in the data set in year $(t + 1)$ with future teacher assignments. We include teacher indicator variables for students' fifth-grade teacher assignments and test the null hypothesis that these different teacher assignments differentially

13. In fact, a test for the statistical significance of the student fixed effects in equation 6 fails to reject the null hypothesis of joint insignificance (tested using the gain score analog to this model). However, the test is of low power given the large-N, small-T panel data set structure (typical of most value-added analyses), limiting inference.
14. Rothstein (2010) takes an entirely different approach based on Chamberlain's correlated random effects model when testing for student fixed effects in his analysis.
15. Although all three of these studies use the same basic methodology, Harris and Sass (2006) estimate their model using the generalized method of moments, while Koedel (2009) and Koedel and Betts (2007) use two-stage least squares (2SLS). We use 2SLS here.
16. We include lagged teacher assignments for all lagged teachers who teach at least five students in our sample in the prior year.

predict grade 4 test score growth.[17] So adjusting model 4 by adding controls for grade 5 teachers,

$$Y_{it}^4 = \psi_t + Y_{i(t-1)}\phi_1 + X_{it}\phi_2 + T_{it}^4\theta + T_{i(t+1)}^5\eta + \varepsilon_{it} \tag{4'}$$

where η is a vector of future (grade 5) teacher fixed effect coefficients. We test whether the indicators for future teachers differentially predict achievement in the current year: $H_0 : \eta_1 = \eta_2 = \ldots = \eta_J = \bar{\eta}$. A rejection of this null hypothesis for future teacher effects suggests that sorting bias is contaminating the teacher effects in the model. This is the falsification test proposed by Rothstein (2010).

It is less straightforward to add the future teacher effects to the student fixed effects model because of the first-differencing procedure. For example, a student's fourth-grade teacher enters into the model for third-grade value added as a future teacher and into the model for fourth-grade value added as a current teacher. We allow fourth-grade teachers to have one effect in the lagged score model and a separate effect in the current score model by not differencing out the teacher indicator variables. This approach is taken because the current score teacher effect may be partially causal while the future teacher effect cannot be. The current score and lagged score effects are not separately identifiable, but they are captured by a single coefficient for each fourth-grade teacher. Equation 7 details the first-differenced version of the within-students model that incorporates future teacher assignments. Year t corresponds to the fourth grade for the students in our sample.

$$\begin{aligned}
Y_{it} &= \alpha_i + \lambda_t + Y_{i(t-1)}\tau_1 + X_{it}\tau_2 + S_{it}\tau_3 + T_{it}^4\delta_4 + T_{i(t+1)}^5\eta_5 + u_{it} \\
Y_{i(t-1)} &= \alpha_i + \lambda_{(t-1)} + Y_{i(t-2)}\tau_1 + X_{i(t-1)}\tau_2 + S_{i(t-1)}\tau_3 + T_{i(t-1)}^3\delta_3 \\
&\quad + T_{it}^4\eta_4 + u_{i(t-1)} \\
Y_{it} - Y_{i(t-1)} &= (\alpha_i - \alpha_i) + (\lambda_t - \lambda_{(t-1)}) + (Y_{i(t-1)} - Y_{i(t-2)})\tau_1 \\
&\quad + (X_{it} - X_{i(t-1)})\tau_2 + (S_{it} - S_{i(t-1)})\tau_3 + T_{it}^4\delta_4 - T_{i(t-1)}^3\delta_3 \\
&\quad + T_{i(t+1)}^5\eta_5 - T_{it}^4\eta_4 + (u_{it} - u_{i(t-1)})
\end{aligned} \tag{7}$$

In equation 7 we instrument for the lagged test score gain with the second-lagged test score level. The second row in the equation contains the vectors of teacher effects after first-differencing. The positive entries are from the current

17. By not requiring all students to have future teacher assignments we are able to use a larger student sample, and therefore a larger teacher sample. The reference group here is the student population for which no grade 5 teacher is observed. The results do not depend on the use of this comparison group. For instance, dropping the students without a fifth-grade teacher and instead omitting one future teacher does not change the results.

score model and the negative entries are from the lagged score model. The superscripts on the teacher indicator vectors indicate the grade level taught by the teachers, along with the corresponding subscripts on the coefficient vectors.[18] The teacher coefficients denoted by δ may contain some causal component, while the coefficients denoted by η cannot possibly contain causal information. Grouping terms, the vector of current teacher coefficients in this model estimates $(\delta_4 - \eta_4)$.

In each model, we estimate future teacher effects for all teachers who teach at least twenty students from our original student sample, one year in advance. We perform Wald tests of the null hypothesis that the teacher effects are jointly equal to each other separately for current and future teachers, although our primary interest is in the tests for the future teachers. We also estimate the unadjusted and adjusted variances of the distributions of the current and future teacher effects, again following Aaronson, Barrow, and Sander (2007).[19]

A caveat to the falsification tests in the basic and within-schools models is that they need not indicate sorting bias. Returning to equation (4′) and the analogous within-schools model, if students are sorted into year (t + 1) classrooms based on year t performance, grade 5 teacher assignments *should* be correlated with grade 4 test score growth. By itself, this is not an indictment of the VAM because a model of grade 5 test score growth, which would be used to estimate grade 5 teacher effects, would include grade 4 performance as a control. However, only if grade 4 test scores fully capture differences in academic readiness prior to grade 5 would this benign explanation for grade 5 teacher effects suffice. If this condition fails, the future teacher coefficients in the basic and within-schools models will capture sorting bias attributable to any unobserved factors that determine both academic performance and student-teacher assignments. In the context of the models, if the error terms in the basic and within-schools models are serially correlated, nonzero future teacher effects will be problematic.[20]

One way to test the extent to which the future teacher effects in these models are capturing bias from sorting on unobservables is to look at prior year

18. Our approach requires that we treat teacher effects separately by grade for fourth-grade teachers who also teach students in the third grade. If teacher effects are constant across grades, these by-grade effects are expected to difference out for a student who has the same teacher in the third and fourth grades (assuming constant quality). However, this creates some additional noise relative to a standard first-differenced model because more than one parameter must be estimated for the forty-nine fourth-grade teachers who also teach in the third grade in our panel.
19. We diagonalize the variance matrices to compute the Wald statistics. Substituting the full variance-covariance matrices for the diagonal variance matrices has little effect on the reported Wald statistics, and mechanically it does not affect the teacher effect variance estimates at all.
20. We expect that the student fixed effects model will reduce the problem of serial correlation because in the simpler models the errors will partly reflect omitted student ability and motivation, which are largely unchanging across grades.

teachers. If lagged test scores are complete measures of academic readiness, which is the condition required for the future teacher effects to be benign, the contributions of prior year teachers to student performance should be small. Rothstein (2010) shows that this is not the case, and in fact lagged teacher assignments are strong predictors of student outcomes in VAMs.[21] This suggests that nonzero future teacher effects in the basic and within-schools models will reflect sorting bias.

Turning to the within-students model, the falsification test is a test of strict exogeneity. By assumption, the time-varying information contained in year t test scores does not influence teacher assignments in year (t + 1) in this model. Nonzero future teacher effects indicate that the future year teacher assignments are correlated with the first-differenced error terms—that is, they indicate that strict exogeneity fails.

Table 4 details our initial results from the three VAMs—in all three specifications, the significance of the future teacher effects cannot be rejected. Although the results in table 4 continue to show nonzero future teacher effects, note that the estimates of the standard deviations of teacher effects are much smaller than the analogous estimates in table 1. For example, the ratio of the adjusted standard deviation of the future teacher effects distribution to the adjusted standard deviation of the current teacher effects distribution falls below one-half in each of the models in table 4 (down from approximately 0.6 in table 1).[22] Compared with table 1, table 4 indicates that richer VAMs that evaluate teachers over multiple years reduce the bias in the estimated teacher effects.[23]

One potentially important aspect of the results in table 4 is that some of the future teacher effects are estimated using multiple cohorts of students. If the sorting captured by Rothstein's estimates (and our analogous estimates) is transitory to some extent, using multiple cohorts of students to evaluate teacher effects will help mitigate the bias. To illustrate, we write the single-year teacher effect estimate for teacher j at school k in year t from the basic VAM as the sum of five components:

$$\hat{\theta}_{jkt} = \theta_j + \lambda_k + (\delta_j + \rho_{jt}) + \nu_{jkt}. \tag{8}$$

21. We replicate this finding in our data (results available upon request).
22. Note that the meaning of this ratio is less clear in the student fixed effects model because the current teacher effects from this model estimate the joint parameter $(\delta_4 - \eta_4)$ in equation 7. Ultimately, however, the important result from the student fixed effects model is that the future teacher effects have a less predictive power over current test score growth.
23. The control variables added to the specifications marginally reduce the sorting bias. This result is consistent with Rothstein (2010). Although Rothstein does not report results from models that incorporate student- or school-level control variables, he notes that his results do not qualitatively change if they are included in the model.

Table 4. Extension of Rothstein's Analysis Using the Value-Added Models from Section 5

		Wald Statistic (DF)	*P*-Value	Unadjusted Standard Deviation	Adjusted Standard Deviation
Basic model	Grade 4 teachers	3393 (594)	<0.01	0.27	0.23
	Grade 5 teachers	846 (471)	<0.01	0.16	0.10
Within-schools model	Grade 4 teachers	1994 (594)	<0.01	0.28	0.22
	Grade 5 teachers	815 (471)	<0.01	0.15	0.10
Within-students model	Grade 4 teachers[a]	649 (388)	<0.01	0.29	0.18
	Grade 5 teachers	341 (259)	<0.01	0.16	0.07

See notes to table 1.
[a]Note that these estimates are for the composite effects documented in equation 7.

In equation 8, θ_j is teacher j's true effect, λ_k measures the quality of school k, δ_j measures bias from persistent student sorting to teacher j across years, ρ_{jt} measures bias from nonpersistent student sorting to teacher j, and v_{jkt} is the statistical noise associated with the teacher effect estimate. If $\lambda_k \neq 0$, the within-schools model is appropriate. Including additional student-level controls and/or student fixed effects will reduce the impacts of δ_j and ρ_{jt}; however, if students are sorted to teachers based on dynamic and unobserved attributes that are correlated with test score growth (e.g., expected mean reversion), these terms can be nonzero in expectation even in the within-students model.

Using multiple years of data to evaluate teachers will reduce the bias from sorting on unobservables to the extent that the sorting is transitory and captured by ρ_{jt} rather than δ_j. As long as student sorting is partly transitory, the variance of our estimated teacher effects will fall with the number of cohorts observed for each teacher j because ρ_{jt} is being averaged over an increasing number of cohorts of students. In the extreme, if sorting on unobservables contributes only to ρ_{jt} (i.e., it is entirely nonpersistent from year to year), the sorting bias will go to zero as t increases.

Table 5 provides compelling evidence that transitory sorting bias may be an important concern in the data. We group students by year (t + 1) classrooms, then calculate the within-teacher, across-year correlations of classroom average year t gain scores for the year (t + 1) teachers who teach in each year of the within-students data panel. In the context of equation 8, for teacher j these correlations provide a rough measure of the relationship between $(\delta_j + \rho_{jt})$ and $(\delta_j + \rho_{j(t+1)})$. Correlations near one would suggest that there is little scope for transitory bias and that adding additional years of teacher data will not be helpful. Correlations near zero would indicate that the sorting bias can be reduced by evaluating teachers across multiple years. The first panel of table 5 reports correlations for raw year t gain scores, which range between 0.30

Table 5. Across-Year Correlations in Year t Gain Scores Averaged at the Teacher-by-Year Level for Year (t + 1) Teachers Who Taught in All Three Years of the Within-Students Panel

	Year 1	Year 2	Year 3
Raw Gain Scores			
Year 1	1		
Year 2	0.35	1	
Year 3	0.30	0.34	1
Year t Gain Scores Demeaned within Students			
Year 1	1		
Year 2	0.20	1	
Year 3	0.14	0.30	1

and 0.36 across years. The second panel reports correlations for year t gain scores that are demeaned within students (i.e., we subtract students' average gains), which are even smaller and range from 0.14 to 0.30. Although these correlations are not zero, they are far from one, suggesting that transitory sorting bias can be reduced by evaluating teachers across multiple years.

As a second way to investigate the significance of transitory sorting bias, we replicate our analysis from table 4 but evaluate only future teachers who teach students in every possible year of the data panel. For the basic and within-schools models, this means that future teachers teach students in four consecutive years. For the within-students model, future teachers teach students in three consecutive years (recall that we use only three year-cohorts of students in the within-students model).

We report our results in table 6. Consistent with what is suggested by the correlations in table 5, future teacher effects are smaller when we focus on future teachers who teach multiple cohorts of students. In fact, in the student fixed effects model, when we focus on future teachers who teach at least three classrooms of students, the adjusted variance of grade 5 teacher effects goes to zero. This suggests that at least some of the sorting bias uncovered by Rothstein (2010) is transitory.[24] This finding highlights perhaps the most policy-relevant implication of our study—evaluating teachers over multiple years will improve the performance of VAMs and, depending on the sorting environment, may

24. Our transitory sorting bias finding is consistent with other work that finds that multiyear teacher effects are more stable (McCaffrey et al. 2009) and more predictable (Goldhaber and Hansen 2008). However, reduced sampling variance will also be a determinant of these other results.

Table 6. Extension of Rothstein's Analysis Using the Value-Added Models from Section 5 and Modeling Only Future Teachers Who Taught Students in Each Year of the Data Panel

		Wald Statistic (DF)	*P*-Value	Unadjusted Variance (s.d.)	Adjusted Variance (s.d.)
Basic model	Grade 4 teachers	4640 (594)	<0.01	0.27	0.23
	Grade 5 teachers	268 (140)	<0.01	0.14	0.09
Within-schools model	Grade 4 teachers	2260 (594)	<0.01	0.28	0.23
	Grade 5 teachers	260 (140)	<0.01	0.14	0.08
Within-students model	Grade 4 teachers[a]	684 (388)	<0.01	0.29	0.19
	Grade 5 teachers	158 (147)	0.25	0.13	0.00[b]

Notes: See notes to table 1. For the basic and within-schools models, this analysis includes fifth-grade teachers who teach in all four years of our data panel. In the within-students model we evaluate just three year cohorts of students, and therefore we include fifth-grade teachers who teach in three consecutive years.
[a]Note that these estimates are for the composite effects documented in equation 7.
[b]Adjusted variance estimate was marginally negative.

be sufficient to mitigate sorting bias if static tracking is adequately controlled for.[25]

6. IS THIS TRANSITORY SORTING BIAS OR SAMPLE SELECTION?

In addition to transitory sorting bias, sample selection may also partly explain our results in table 6. For example, by requiring teachers to teach in all three years of our data panel, we exclude a disproportionate share of inexperienced teachers. If students are sorted differently to experienced and inexperienced teachers, this could contribute to our findings. Although we cannot capture fully the differences between the teachers who do and do not exit the data panel, we can replicate the correlative analysis in table 5 separately for the experienced and inexperienced teachers who remained in the data set for all three years. We present these correlations in table 7—they do not suggest that the magnitude of transitory sorting bias will differ across experienced and inexperienced teachers.[26]

We also directly test the extent to which our results in table 6 are driven by sample selection. If what we have uncovered is a sample selection effect, if we remove a cohort of student data and rerun the model using the new student subsample but the same teachers, the adjusted variance of grade 5 teacher effects should remain near zero, and the Wald test should continue

25. Although we focus on future teachers, the analysis is also relevant for lagged teachers. In omitted results we find that the patterns of bias associated with transitory sorting for future teachers are also reflected among lagged teachers, although the interpretation of the lagged teacher analysis is less straightforward for a handful of reasons, some specific to our data set (contact the authors for details).
26. We omit the demeaned gain score correlations for brevity—they are smaller in magnitude and display a similar pattern to the raw gain scores.

Table 7. Across-Year Correlations in Year t Gain Scores Averaged at the Teacher-by-Year Level for Year (t + 1) Teachers Who Taught in All Three Years of the Within-Students Panel, by Teacher Experience

	Year 1	Year 2	Year 3
Raw Gain Scores: Experienced Teachers ($N = 104$)			
Year 1	1		
Year 2	0.27	1	
Year 3	0.33	0.38	1
Raw Gain Scores: Inexperienced Teachers ($N = 43$)			
Year 1	1		
Year 2	0.59	1	
Year 3	0.22	0.23	1

Table 8. Within-Students Model Using Only Future Teachers Who Taught Students in Each Year of the Data Panel, with Each of the Three Year Cohorts Individually Omitted from the Data Set

		Wald Statistic (DF)[a]	*P*-Value	Unadjusted Variance (s.d.)	Adjusted Variance (s.d.)
Drop fourth-grade cohort in 1999–2000	Grade 4 teachers[b]	568 (369)	<0.01	0.31	0.18
	Grade 5 teachers	181 (147)	0.03	0.17	0.06
Drop fourth-grade cohort in 2000–1	Grade 4 teachers[b]	651 (374)	<0.01	0.34	0.21
	Grade 5 teachers	190 (147)	0.01	0.20	0.09
Drop fourth-grade cohort in 2001–2	Grade 4 teachers[b]	617 (332)	<0.01	0.36	0.23
	Grade 5 teachers	196 (147)	<0.01	0.21	0.07

See notes to table 1.
[a]The number of grade 4 teachers included in the model changes across rows because some of the grade 4 teachers taught only in a single year. Note that the 2001–2 cohort of students was somewhat larger than the other two cohorts, which explains why there are fewer grade 4 teachers in the model when this cohort is dropped.
[b]These estimates are for the composite effects documented in equation 7.

to retain the null that all grade 5 teachers have an identical effect on grade 4 achievement. Table 8 shows the results when we reestimate the within-students model after removing one cohort of grade 4 students at a time. The adjusted variance of grade 5 teachers now rises. Also, in two of three cases the adjusted variance for the grade 4 teachers also increases as would be expected.

So why does the fixed effects model using future teachers who teach in all years, shown in table 6, appear to salvage hope for the use of VAMs? We conclude that there is not something unusual about this sample of grade 5 teachers. Rather, the main reason we succeed in reducing future teacher effects to zero has mostly to do with the fact that in table 6 we include only

grade 5 teachers who teach in all years in the data. The use of multiple years of data reduces transitory sorting bias significantly.[27]

7. THE VARIANCE OF TEACHER QUALITY IN SAN DIEGO

The results from the previous section suggest that we can estimate the variance of causal teacher effects in San Diego using a within-students VAM that focuses on teachers who teach in all three years of our data panel. For this analysis we return to the within-students model in equation 6 from section 5 and estimate teacher effects for fourth-grade teachers. Unlike in the previous analysis, we do not include future teachers in the model, and we estimate a typical first-differenced specification (as opposed to the nonstandard specification in equation 7).

Across all the fourth-grade teachers in our within-students sample, the adjusted standard deviation of the teacher effects from the model in equation 6 is estimated to be 0.22; this number is similar in magnitude to the results above.[28] To estimate the magnitude of the variance of actual teacher quality, free from sorting bias, we split the teacher sample into two groups. Group A consists of fourth-grade teachers who taught in all three years of our within-students data panel, and group B consists of teachers who did not. Approximately 45 percent of the fourth-grade teachers belong to group A and 55 percent belong to group B.[29] Consistent with the transitory sorting bias result in table 6, the adjusted variance of the teacher effects from group A is approximately 24 percent smaller than the adjusted variance of the teacher effects from group B. Correspondingly, the standard deviations of the adjusted teacher effect distributions, measured in standard deviations of the test, are 0.20 for group A and 0.23 for group B. The standard deviation of the adjusted difference in variance between the two groups is 0.11. Table 9 documents these results.

Although the analysis in the previous section suggests that the observed variance gap between the teachers in groups A and B will be driven, at least in

27. In an analysis omitted for brevity, we investigated the extent to which transitory student sorting is exacerbated by principal turnover. Although our findings are consistent with a principal-turnover effect in the expected direction, the effect is not statistically significant. One implication of this result is that transitory student sorting does not unduly depend on principal turnover. Further details are available from the authors upon request.

28. Recall that the within-student teacher effect estimates in tables 6 and 8 are from the nonstandard first-differenced model in equation 7.

29. Note that in table 6, roughly 57 percent of fifth-grade teachers taught in all three years of the data panel. The difference in stability between our fourth- and fifth-grade teacher samples may be explained by the different selection criteria. Our initial sample of fifth-grade teachers in table 4 teach at least twenty students for whom we observe teacher assignments in four consecutive years, while our sample of fourth-grade teachers teach at least twenty students for whom we observe teacher assignments in just three consecutive years. In addition, the fifth-grade teacher sample is identified conditional on students being taught by one of the teachers in the fourth-grade teacher sample.

Table 9. Teacher-Effect Variance Estimates from the Within-Students Model (Equation 6)

	Unadjusted Variance[a]	Adjusted Variance[a]	Standard Deviation (adjusted)[b]
All teachers	128	76	0.22
Group A	110	64	0.20
Group B	141	84	0.23
Variance gap	31	20	0.11

Notes: Teachers in group A taught in all three years of the data panel; teachers in group B did not.
[a]The unadjusted and adjusted variance estimates are in raw test score points.
[b]The standard deviation estimates are ratios of the standard deviations of the teacher effect (and bias) distributions to the standard deviation of the fourth-grade test score distribution. These estimates are analogous to those presented in tables 2, 5, and 6.

Table 10. Differences in Observable Characteristics between Teachers Who Taught in All Three Years of Our Data Panel (Group A) and Those Who Did Not (Group B)

	Group A	Group B
Share with experience of less than 5 years	0.27	0.48
Share with experience between 5 and 10 years	0.20	0.19
Share with experience of more than 10 years	0.53	0.33
Share with MA degree	0.65	0.48
Share with BA in education	0.40	0.38

Note: Characteristics are averaged within teachers over the course of the data panel.

part, by differences in transitory sorting bias, two other explanations merit discussion. First, again, sample selection may be a concern if group A is a more homogeneous group of teachers than group B. As shown in table 10, there are some observable differences in experience and education that suggest this might be a concern. Specifically, teachers in group A are likely to be more experienced and to have a master's degree. We investigate the extent to which differences across groups along these dimensions might explain the observed variance difference by estimating the within-group variance of teacher quality for more and less experienced teachers, and then for teachers with and without master's degrees. The within-group variance of teacher effects among teachers with master's degrees is higher than the within-group variance of those without, which works counter to the observed variance gap. For experience, there is more variation among teachers with ten or more years of experience and among novice teachers (with five or fewer years of experience) than among teachers with five to ten years of experience. Ultimately, the variance

decompositions based on grouping teachers by observable qualifications do not suggest a clear variance gap effect.[30]

We also note that the grouping criterion here is somewhat arbitrary in the sense that there is nothing particularly special about the years covered by our data panel. For example, some of the teachers in group A surely left the district in the year after our data panel ended or did not teach in the year before it started, and some of the teachers in group B surely taught in three or more contiguous years outside the data panel (for example, if a teacher taught in the year prior to the first year of our data panel, and then in the first two years of our data panel but not the third, we would assign the teacher to group B).

The second explanation for the observed variance gap is that it occurs by chance. To evaluate this possibility, we use a bootstrap to derive empirically the distribution from which the variance gap estimate would be drawn if the sample were split at random. We randomly assign the teachers from our sample into two groups that are equivalent in size to groups A and B above, and calculate the adjusted variance gap between these randomly assigned groups. We repeat this procedure five hundred times and use the five hundred variance gap estimates to define the variance gap distribution. The variance gap is calculated as the adjusted variance of teacher effects in the smaller group minus the adjusted variance of teacher effects in the larger group, all divided by the adjusted variance of teacher effects in the larger group. In other words, we calculate $[var(Group\,A) - var(Group\,B)]/var(Group\,B))$. The average variance gap generated by the bootstrap analysis is +1 percent. The standard deviation in this variance gap is quite large, though, at 24 percent. Thus the variance gap estimated between the teachers in groups A and B (−24 percent as shown in the second column of table 9) is just over a standard deviation away from the average of the empirical variance gap distribution (at approximately the 13th percentile of the range of bootstrapped estimates). Although the empirical variance gap distribution is wide (the 90 percent confidence interval ranges from −35 to +45 percent), which limits our ability to detect statistical significance even when the observed variance gap is large, the gap estimated between groups A and B is suggestive of a transitory sorting bias effect.

30. The differences in variances across the teacher samples split by observable qualifications are small, in the neighborhood of 0.01–0.02 standard deviations. Although we cannot disentangle the effects of transitory sorting bias from the observable differences across teachers in the two samples of interest (groups A and B), there is a large literature showing that teachers differ only mildly in effectiveness based on observable qualifications (Hanushek 1996; exceptions in the literature include Clotfelter, Ladd, and Vigdor 2007). Perhaps most relevant to the present study, Betts, Zau, and Rice (2003) estimate VAMs in the SDUSD using student fixed effects, with separate models for elementary, middle, and high school students. Although they find some evidence that teacher qualifications matter at the high school level, they find very little evidence of this in elementary schools.

8. CONCLUSION

On the one hand, our results corroborate Rothstein's key finding that VAMs of student achievement can produce biased estimates of teacher effects. In fact, we show that even detailed VAMs that estimate teacher effects across multiple cohorts of students can still produce biased estimates, as evidenced by the future teacher effects documented in tables 4, 6, and 8. On the other hand, our results are encouraging because they indicate that sorting bias in value-added estimation need not be as large as is implied by Rothstein's work. A key finding here is that using multiple years of classroom observations for teachers will reduce sorting bias in value-added estimates. This result raises concerns about using single-year measures of teacher value added to evaluate teacher effectiveness. For example, one may not want to use achievement gains of the students of novice teachers who are in their first year of teaching to make decisions about which novice teachers should be retained.

In our setting in San Diego, using a student fixed effects model and evaluating teachers who teach students in three consecutive years mitigates the contribution of sorting bias to the teacher effect estimates. Although this result may not universally generalize and depends on the degree of student-teacher sorting in our data, it suggests that under some circumstances value-added modeling can continue to be a powerful tool in the analysis of teacher effectiveness.

Nonetheless, to the extent that our results corroborate Rothstein's findings, they highlight an important issue with incorporating value-added measures of teacher effectiveness into high-stakes teacher evaluations. Namely, value added is manipulatable by administrators who determine students' classroom assignments. Our entire analysis is based on a low-stakes measure of teacher effectiveness. If high stakes were assigned to value-added measures of teacher effectiveness, sufficient safeguards would need to be put in place to ensure that the system could not be gamed through purposeful sorting of students to teachers for the benefit of altering value-added measures of teacher effectiveness.

The authors thank Andrew Zau and many administrators at the San Diego Unified School District (SDUSD), in particular Karen Bachofer and Peter Bell, for helpful conversations and assistance with data issues. We also thank Zack Miller, Shawn Ni, Mike Podgursky, seminar participants at Northwestern University and Simon Fraser University, and two anonymous referees for useful comments and suggestions, and the National Center for Performance Incentives for research support. SDUSD does not have an achievement-based merit pay program, nor does it use value-added student achievement data to evaluate teacher effectiveness. The underlying project that provided the data for this study has been funded by a number of organizations, including the William and Flora Hewlett Foundation, the Public Policy Institute of California, the Bill

and Melinda Gates Foundation, the Atlantic Philanthropies, and the Girard Foundation. None of these entities has funded the specific research described here, but we warmly acknowledge their contributions to the work needed to create the database underlying the research.

REFERENCES

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

Anderson T. W., and Cheng Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76: 598–609.

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

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41: 778–820.

Goldhaber, Dan, and Michael Hansen. 2008. Is it just a bad class? Assessing the stability of measured teacher performance. CRPE Working Paper No. 2008-5.

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

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin. 2005. The market for teacher quality. NBER Working Paper No. 11154.

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

Harris, Douglas, and Tim R. Sass. 2007. What makes for a good teacher and who can tell? Unpublished paper, Florida State University.

Jacob, Brian, and Lars Lefgren. 2007. Principals as agents: Subjective performance assessment in education. NBER Working Paper No. 11463.

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

Kane, Thomas, and Douglas Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Koedel, Cory. 2009. An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review* 28(6): 682–92.

Koedel, Cory, and Julian R. Betts. 2007. Re-examining the role of teacher quality in the educational production function. Working Paper No. 07–08, University of Missouri, Columbia.

Koedel, Cory, and Julian R. Betts. 2010. Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy* 5(1): 54–81.

McCaffrey, Daniel F., J. R. Lockwood, Tim R. Sass, and Kata Mihaly. 2009. The intertemporal variability of teacher effect estimates. *Education Finance and Policy* 4(4): 573–607.

Murnane, Richard J., Judith D. Singer, John B. Willett, James J. Kemple, and Randall J. Olson. 1991. *Who will teach? Policies that matter*. Cambridge, MA: Harvard University Press.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26: 237–57.

Podgursky, Michael J., and Mathew G. Springer. 2007. Teacher performance pay: A survey. *Journal of Policy Analysis and Management* 26: 909–50.

Rockoff, Jonah. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94(2): 247–52.

Rothstein, Jesse. 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy* 4(4): 538–72.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125(1): 175–214.