USC Dornsife
Center for Economic and Social Research

USC Schaeffer
Leonard D. Schaeffer Center for Health Policy & Economics

Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition?

Gema Zamarro, John Engberg,
Juan Esteban Saavedra, Jennifer Steele

Paper No: 2014-001

CESR-SCHAEFFER
WORKING PAPER SERIES

The Working Papers in this series have not undergone peer review or been edited by USC. The series is intended to make results of CESR and Schaeffer Center research widely available, in preliminary form, to encourage discussion and input from the research community before publication in a formal, peer-reviewed journal. CESR-Schaeffer working papers can be cited without permission of the author so long as the source is clearly referred to as a CESR-Schaeffer working paper.

cesr.usc.edu **healthpolicy.usc.edu**

Disentangling Disadvantage: Can We Distinguish Good Teaching from Classroom Composition?

By GEMA ZAMARRO, JOHN ENGBERG, JUAN ESTEBAN SAAVEDRA AND JENNIFER STEELE*

Abstract

This paper focuses on the use of teacher value-added estimates to assess the distribution of effective teaching across students of varying socioeconomic disadvantage. We use simulation methods to examine the extent to which different commonly used teacher-value added estimators accurately capture both the rank correlation between true and estimated teacher effects and the distribution of effective teaching across student characteristics in the presence of classroom composition effects. Varying the amount of teacher sorting by student characteristics, the within-teacher variability in classroom composition, and the amount of student learning decay, we compare aggregated residuals, teacher random effects, and teacher fixed effects models estimated in both levels and gains, with and without controls for classroom composition. We find that models estimated in levels more accurately capture the rank correlation between true and estimated teacher effects than models estimated in gains, but levels are not always preferable for recovering the correlation between teacher value-added and student achievement. For recovering that correlation, aggregated residuals models appear preferable when sorting is not present, though fixed effects models perform better in the presence of sorting. Because the true amount of sorting is never known, we recommend that analysts incorporate contextual information into their decisions about model choice.

JEL codes: I24, C01, I29
Keywords: value-added models, Monte Carlo simulation, teacher quality, classroom composition

* Corresponding author Zamarro: University of Southern California Dornsife Center for Economic and Social Research & RAND Corporation, 12015 Waterfront Drive, Playa Vista CA 90094 (email: zamarror@usc.edu); Engberg: RAND Corporation, 4570 Fifth Avenue, Suite 600 Pittsburgh, PA 15213-2665, (email: engberg@rand.org); Saavedra: University of Southern California Dornsife Center for Economic and Social Research, 12015 Waterfront Drive, Playa Vista CA 90094 (email: juansaav@usc.edu); Steele: RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202-5050 (email: jsteele@rand.org). **Acknowledgements:** The authors thank Manuel Arellano, Erik Meijer, Jonah Rockoff, James Wyckoff, Brian Stecher, and seminar participants at AEFP 37th Annual Conference, IES funded Conference "Using Student Test Scores to Measure Teacher Performance: The State of the Art in Research and Practice" , and CEMFI. for helpful comments and suggestions. We also wish to thank the Bill & Melinda Gates Foundation and the USC Dornsife Center for Economic and Social Research, which provided funding for this research. The opinions expressed here are solely those of the authors and do not represent those of the RAND Corporation, the University of Southern California or any of its sponsors. Any errors are the responsibility of the authors.

1. Introduction

The unequal access to effective teaching is a long-standing policy concern. High-poverty schools, for example, tend to concentrate higher proportions of novice and academically weak teachers. (Becker, 1952; Clotfelter, Ladd, & Vigdor, 2005; Lankford, Loeb, & Wyckoff, 2002). High-poverty schools also face disproportionate difficulties in retaining teachers, who often leave to teach in more affluent schools (Feng, 2010; Hanushek, Kain & Rivkin, 2004). Motivated by the challenges that high-poverty schools face in attracting and retaining qualified teachers, recent public and private efforts to close the achievement gap between disadvantaged and more advantaged students increasingly focus on redistributing effective teachers.

The No Child Left Behind Act (NCLB) requires, for example, that states improve the equitable distribution of highly qualified teachers as measured by their licensure and subject matter preparation. Race to the Top—another federal government initiative—rewards states that convincingly plan to improve the share of highly effective teachers in high-poverty schools. Meanwhile, the Bill & Melinda Gates Foundation is collaborating with a group of districts and charter management organizations across the country to refine their human resource policies in ways that improve disadvantaged students' access to effective teachers. Similarly, the U.S Department of Education, Institute of Education Sciences (IES) funded Talent Transfer Initiative, offered a financial incentive to the teachers with the highest scores on value-added measures if they would transfer to a lower-achieving school in the same district and remain there for at least two years.

However, teacher attributes traditionally used to proxy for teacher effectiveness—academic credentials, licensure and subject matter preparation, for example—are only weakly connected to teachers' effects on student achievement (Aaronson, Barrow, & Sander, 2007; Kane

& Staiger, 2005; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). For this reason, the focus of current efforts that aim to make the distribution of teacher effectiveness more equitable has shifted from redistributing effective teachers based on their observable attributes, to redistributing effective teachers according to their effect on student achievement, commonly referred to as their value-added.

Teacher value-added is usually understood as the contribution of a teacher to the learning of his/her students as measured by test scores and conditional on other factors, including students' background and preparation. Teachers are then considered to be highly effective if their students make above average improvements relative to other teachers with comparable students (see, for example, Buddin & Zamarro, 2009; Hannaway et al., 2010).

Estimating a teacher's impact on student learning requires, however, that other inputs that affect student learning be appropriately controlled for. Typically, past inputs are captured through measures of prior student learning such as lagged test-scores, which may be inaccurate for a number of reasons. Lagged test-scores may fail to capture, for example, all baseline knowledge relevant to the focal year's test. They may also fail to capture all features of innate ability, motivation, and family circumstances that affect student learning. Lagged test scores also contain measurement error and do not reflect an unknown amount of learning decay.

In addition to past inputs to learning, teacher value-added models also aim to control for educational inputs that are contemporaneous with the teacher's contribution. Some contemporaneous inputs are intrinsic to the student and are imperfectly measured by his or her own characteristics such as gender, race and free lunch status. Other contemporaneous inputs include classroom composition characteristics, which may influence a student's experience in a classroom independently of the teacher who is leading the class, and which may also be imperfectly measured by observable classroom averages of gender, race and free lunch status, among other variables. For example, a large class or the presence of a few extremely difficult students who require special attention may make it harder for other students to learn.

For these reasons, teacher value-added models are an imperfect approach to identifying teachers' impact on student achievement. Moreover, models that estimate teacher's value-added most precisely—i.e., that minimize mean squared error—may not be the models that best recover unbiased estimates of a distribution parameter that measures the correlation between teacher effectiveness and student characteristics. Some teacher value-added models may yield biased estimates of the correlation between a student's demographic characteristics and teacher's effectiveness due to unobserved student sorting and model specification choices.

Chetty, Friedman, and Rockoff (2011) argue, for example, that teacher value-added estimates based on aggregated residuals—in which teacher effects are estimated from unexplained variance after accounting for student and classroom features—are unbiased estimates of teachers' contributions to student achievement. Hannaway and colleagues (2010), on the other hand, argue that models that in addition to including teacher fixed effects also include student fixed effects, are more suitable to estimating the distribution of effective teaching among students of varying socioeconomic disadvantage. Using administrative data from a large school district in the U.S., we demonstrate that the choice of teacher value-added model affects estimates of the distribution of effective teaching among students of varying socioeconomic disadvantage. This finding partly motivates our simulation approach.

While attempts have been made at quantifying the bias of different teacher value-added models to measure teacher effectiveness using simulation methods (See for example, Rothstein, 2009 or Guarino, Reckase, and Wooldridge, 2013), there is no previous effort at quantifying the bias in estimates of a distribution parameter that measures the correlation between teacher effectiveness and students' socioeconomic background obtained from various teacher value-added models. This paper aims to address this outstanding research gap.

To understand the suitability of various teacher value-added models for estimating the relationship between teacher effectiveness and students' socioeconomic characteristics, this paper employs a simulation approach in which we vary our model specifications as well as our assumptions about student sorting and classroom composition effects. Specifically, we address three research questions:

1. Which of several popular value-added modeling approaches (including the aggregate residuals model, teacher random effects, and teacher fixed effects models most accurately captures the (a) rank correlation between true and estimated teacher effects and (b) correlation between teacher effects and demographic characteristics of their students under varied amounts of teacher sorting when classroom composition effects are present?

2. How sensitive are the results to the amount of student learning decay, to modeling in levels versus gains, and to the amount of within-teacher variation in classroom composition?

3. How does each modeling approach fare with and without the inclusion of controls for classroom-average covariates?

To make these questions more tractable, we make some simplifying choices. Among the possible array of student characteristics, we only consider socioeconomic disadvantage and use teacher value-added estimates to assess the distribution of effective teaching across students of varying socioeconomic disadvantage. Therefore, our data-generating process includes only observed student heterogeneity measured by socioeconomic disadvantage; there is no unobserved heterogeneity in the simulated data. Also, we assume there is no measurement error in lagged test scores.

Our simulation approach begins with the simplest case in which teacher assignment to students is random, which implies that teacher effectiveness is orthogonal to socioeconomic disadvantage at the student and at the classroom level. We then create an alternative scenario in which more-effective teachers are systematically assigned to more socioeconomically advantaged students, so that the percentage of disadvantaged students in a teacher's caseload is negatively correlated with the teacher's effectiveness.

We examine these scenarios in cases with low, medium, and high student learning decay from prior assessments. We also vary the amount of within-teacher variation in classroom composition by allowing high versus low year-to-year mobility of students among schools.

Our study considers three classes of models commonly used in the literature to estimate teachers' value-added effectiveness—aggregated residuals, teacher random effects, and teacher fixed effects. All of these models attempt to mitigate bias in estimates of teacher value-added associated with non-random sorting of students into classrooms and non-random teacher assignment to classrooms. It is not clear, however, how well these models mitigate bias when the parameter of interest is the correlation between teacher effectiveness and students' observed characteristics. For each scenario, we fit models in which the dependent variable is either the current level of student achievement regressed on prior achievement (levels) or the difference between current and lagged achievement (gains). Under each specification, we examine the rank correlation between true and estimated teacher value-added, as well as the true versus estimated correlation between teacher value-added and student characteristics. We also examine the sensitivity of these results to the inclusion or exclusion of classroom characteristics.

The paper is organized as follows: Section 2 empirically motivates our simulation with data from a large, urban district, in which we find that our parameters of interest are very sensitive to modeling decisions. Section 3 describes our assumptions about the education production function, and Section 4 presents our data-generating process for the simulations. Section 5 describes the modeling specifications we choose and their rationales. Section 6 presents results for each scenario and modeling choice, and section 7 concludes with a discussion of implications.

2. The Sensitivity of Estimates of the Distributional Parameter of Teacher Effectiveness

We motivate our simulation exercise with estimates of a distribution parameter that measures the correlation between teacher effectiveness and students' socioeconomic disadvantage based on the three different classes of teacher value-added models described above. The analytic sample includes data for students in grades 2 through 5 for eight consecutive years.

Estimation of the distribution parameter proceeds in three steps. In the first step we estimate a student-level regression in which the dependent variable is students' reading test scores, standardized to have a mean of 0 and a standard deviation of 1 in each year. Covariates in the model include students' one-year lagged reading test scores, a dichotomous student-level OBD indicator representing minority students eligible for subsidized meals, and the proportion of students in the classroom who are OBD, which is intended to separate classroom composition effects from individual teacher effects.

In the second step we “shrink” teacher effect estimates from step one to account for heterogeneity in the amount of information available to estimate each teacher’s contribution to student achievement (Jacob & Lefgren, 2007). In the third step, we estimate the distribution parameter, which is the correlation between shrunken teacher effect estimates and average level of socioeconomic disadvantage (OBD) of all the students ever assigned to that teacher. Table 1 shows estimates of the distribution parameter for the various teacher value added models we consider.

As results in Table 1 show, empirical estimates of the correlation between teacher quality measures and the average level of socioeconomic disadvantage are sensitive to modeling approach, as well as to the inclusion or exclusion of classroom composition controls. Models in levels would conclude a negative correlation while models in gains would conclude a small positive correlation, in all cases except teacher fixed effects with classroom controls. The estimate of the distribution parameter, however, is fairly sensitive to the choice of teacher value-added model employed. For models in levels, teacher random effects (RE) and teacher fixed effects (FE) yield similar distribution parameter estimates of around negative 20 to negative 25 percent, suggesting that in the data, more effective teachers are assigned more often to more socioeconomically advantaged students. Estimates of the distribution parameter based on aggregated residuals (AR) methods in levels suggests, on the other hand, a much smaller correlation between teacher effectiveness and student socioeconomic disadvantage in models of about negative 6 to negative 13 percent. Turning to models estimated with gain scores rather than regression of current on lagged scores, the estimates under all specifications are strikingly smaller, ranging from -0.004 for fixed effects with classroom-level covariates to 0.04 using random effects without classroom-level covariates.

3. **Basic Model for Simulations**

This section describes the theoretical model that underlies our Monte Carlo simulation approach. An education production function is the underlying basis for nearly all recent studies of student achievement. Following Todd and Wolpin (2003), we denote T_{it} as the test score measure of student *i* that is observed in year *t*, and let X_{it} and ξ_{it} represent observed and unobserved inputs for student *i* at time *t*, respectively. We denote by μ_{i0} the student's endowed ability that does not vary over time. We assume that the cognitive production function is linear in the inputs and in the unobserved endowment and that input effects do not depend on the student's age but may depend on the age at which they were applied relative to the current age. Then, a general cognitive production function can be written as:

$$T_{it} = \mu_{i0} + \alpha_1 X_{it} + \alpha_2 X_{it-1} + ... + \delta_1 \xi_{it} + \delta_2 \xi_{it-1} + ... + \omega_{it} \qquad (1)$$

where test scores in a given year are a function of current and past observed and unobserved inputs as well as of the initial ability of the student.

Similarly, the equation for the lagged test score T_{it-1} is:

$$T_{it-1} = \mu_{i0} + \alpha_1 X_{it-1} + \alpha_2 X_{it-2} + ... + \delta_1 \xi_{it-1} + \delta_2 \xi_{it-2} + ... + \omega_{it-1}$$

Multiplying the lagged test score equation by β, the persistence rate of prior learning (where persistence is defined as one minus decay), we get:

$$\beta T_{it-1} = \beta\mu_{i0} + \beta\alpha_1 X_{it-1} + \beta\alpha_2 X_{it-2} + ... + \beta\delta_1 \xi_{it-1} + \beta\delta_2 \xi_{it-2} + ... + \beta\omega_{it-1} \quad (2)$$

Subtracting equations (1) and (2) yields:

$$T_{it} - \beta T_{it-1} = (1-\beta)\mu_{i0} + \alpha_1 X_{it} + (\alpha_2 - \beta\alpha_1) X_{it-1} + \ldots + \\ \delta_1 \xi_{it} + (\delta_2 - \beta\delta_1)\xi_{it-1} + \ldots + (\omega_{it} - \beta\omega_{it-1}) \qquad (3)$$

Assuming constant decay, the value of all prior measured and unmeasured inputs must be decaying at the same constant rate from their time of application, i.e., $\alpha_t = \beta\, \alpha_{t-1}$ and $\delta_t = \beta\delta_{t-1}$, $\forall t$. Then (3) becomes:

$$T_{it} - \beta T_{it-1} = (1-\beta)\mu_{i0} + \alpha_1 X_{it} + \delta_1 \xi_{it} + (\omega_{it} - \beta\omega_{it-1})$$

Adding βT_{it-1} to both sides of the equation we obtain that:

$$T_{it} = (1-\beta)\mu_{i0} + \alpha_1 X_{it} + \beta T_{it-1} + \delta_1 \xi_{it} + (\omega_{it} - \beta\omega_{it-1}) \qquad (4)$$

For simplicity, we assume that the only observed student input is OBD status and that it is time-invariant. Since we can link each student to a specific teacher and classroom assignment, we can represent the effect of these inputs with the time varying parameter γ_{jt}, which denotes the effect of student's classroom (teacher, peers, etc.) at time *t.* Therefore, we operationalize equation (4) with the following two equations:

$$T_{it} = (1-\beta)\mu_{\mathrm{i}0} + \alpha^{'} OBD_i + \beta T_{it-1} + \gamma_{jt} + \upsilon_{it} \qquad (5)$$

$$\gamma_{jt} = \lambda \overline{OBD}_{jt} + \phi_j + \varepsilon_{jt} \qquad (6)$$

In equation 5, time-varying classroom effects are the sum of three components: a permanent teacher effect, ϕ_j, which represents teachers' ability independent of the classrooms they are assigned, a time-varying term represented by the proportion of OBD students the teacher is assigned to ($\overline{OBD}_{jt}$) in a given year, and a random error term, ε_{jt} , which represents other, unmeasured attributes of classroom *j* at time *t* that contribute to the performance of student *i*.

Equations (5) and (6) constitute the basis of our data-generating process. In our simulations, we arbitrarily vary the assignment of teachers to classrooms in a way that creates a correlation between a teacher's effectiveness, ϕ_j , and the average characteristics of her students over all of her classes, $\overline{OBD}_j$ ($\rho = \frac{\mathrm{Cov}(\phi_j, \overline{OBD}_j)}{\sigma_{\phi_j} \cdot \sigma_{\overline{OBD}_j}}$). We refer to this correlation as ρ , and it is our main parameter of interest in the Monte Carlo simulations.

Intuitively, once we control for the direct effect of classroom composition (for $\overline{OBD}_{jt}$) on time-varying classroom effects γ_{jt} , the correlation between the time-invariant teacher effect ϕ_j and the average characteristics of all her students $\overline{OBD}_j$ will be different from zero only if teachers are assigned to classrooms non-randomly, but rather as a function of average classroom composition. Therefore, under random assignment of students to teachers, we would have that ρ =0. When teacher assignment is non-random such that more effective teachers are assigned to classrooms with more socioeconomically advantaged students, $\rho < 0$. Motivated by previous empirical literature on teacher assignment, we only consider cases in which the more effective teachers are assigned to more socioeconomically advantaged students, such that ρ is always less than zero.

4. Data-Generating Process

To make the problem tractable, we make some simplifying choices in our Monte Carlo simulations. As noted, we only consider socioeconomic disadvantage and use the teacher value-added estimates to assess the distribution of effective teaching across students of varying socioeconomic disadvantage. We also ignore possible unobserved student heterogeneity that is not already captured by students' socioeconomic disadvantage. Finally, we assume there is no measurement error in lagged test scores.

Our Monte Carlo simulations use 100 samples of data that we generate from equations (5) and (6). We assume a single cohort of 2,400 students that we observe for four consecutive years. Half of the students in the simulated data are assigned OBD status, and students are assigned into schools randomly, so that the average number of disadvantaged students in each school has a mean of 0.5 and standard deviation 0.3.[1]

We consider a total of 120 teachers so that average classroom size is twenty students per teacher, although in the simulations, class size varies randomly, with a minimum of ten students per teacher. Teachers are assigned to classrooms in a way that produces different values of ρ: random assignment ($\rho \simeq 0$) and noisy sorting on the proportion of OBD students in the classroom ($\rho \simeq -0.5$). For each teacher assignment scenario, we generate 100 simulated samples.

We then consider two scenarios for student and teacher mobility. Our first scenario is such that we have extreme variability in classroom composition. In this case, we ignore the school dimension, and students and teachers are re-assigned into classrooms each year. This scenario corresponds to a case where we have similar levels of between- and within-teacher variation in the data. Our second scenario limits the variability in classroom composition over time, generating a more realistic case where schools are located in certain types of

[1] This distribution replicates the distribution of school-average OBD observed in administrative data from a large, urban school district.

neighborhoods, students remain in the same school for subsequent years, and teachers serve similar types of students year after year. This case results in a larger proportion of between-teacher variation in student characteristics than in the case with high between-school student mobility. This scenario therefore increases the potential noise in teacher fixed effect estimates, as they are estimated using only within-teacher variation in student characteristics. In particular, for this latter scenario, we consider 20 schools with 6 teachers per school. Teachers and students are assigned into schools the first year and not allowed to leave the school in subsequent years. Students are then re-assigned to classrooms each year within schools.

Under each scenario presented above, permanent teacher effects (ϕ_j) are drawn from a normal distribution with mean zero and standard deviation 0.2. Similarly, ε_{jt} in equation (6) above is also a draw from a normal distribution with mean zero and standard deviation 0.1. With these assumptions, we derive time-varying teacher effects from equation (6).[2]

Baseline student test scores are generated following a standard normal distribution, $T_{i0} \sim N(0,1)$. For each subsequent year, student test scores are generated following equation (5) above, assuming υ_{it} is normally distributed with mean and standard deviation such that the resulting test scores have mean zero and variance 1 in each period.

Because we are concerned, especially for models in gains that make the implicit assumption of a zero decay rate, with how the models perform under various scenarios for the persistence of prior learning, we simulate data under different assumptions for the persistence parameter, β $(\beta = 0.8; \beta = 0.4; \beta = 0.2)$ (Note that this parameter represents 1 minus the rate of decay.) The rest of parameter estimates in the equations (5) and (6) are set to correspond to those

[2] The standard deviation of this error distribution is restricted so that the resulting variance of the error in the test score equation υ_{it} has feasible values.

obtained from estimates using data from a large school district. Specifically, these parameters are: $\alpha = -0.08; \lambda = -0.3$, indicating a weak negative correlation between student-level OBD and test scores and a moderately negative correlation between classroom-average OBD and classroom fixed effects.

5. Modeling Specifications Tested

We test three models commonly used to estimate teacher effects: i) aggregated residuals (e.g., Kane and Staiger, 2008); ii) teacher random effects (e.g., McCaffrey et al., 2003); and iii) teacher fixed effects (e.g., Rothstein, 2010) with and without Empirical Bayes adjustment. For each type of model, we fit equations in levels by regressing the current score on explanatory variables that include the lagged score, and we fit equations in gains by omitting the lagged test score from the set of explanatory variables and using gains in test scores as our dependent variable instead. Because levels and gains models utilize different assumptions about the rate of learning decay, with gains presuming no decay and levels estimating the degree of decay (Clotfelter, Ladd, & Vigdor, 2007), it is important to examine the performance of both types of models.

We arrive at our basic estimation equation in levels by substituting (6) into (5) from above:

$$T_{ijt} = \alpha_0 + \alpha_1 T_{it-1} + \alpha_2 OBD_i + \alpha_3 \overline{OBD}_{jt} + \eta_{ijt} \qquad (7)$$

where $\eta_{ijt} = \phi_j + \varepsilon_{jt} + \upsilon_{it}$. In equation (7), T_{ijt} represents the standardized test-score for student i with teacher j in year t, T_{it-1} is student i's lagged test-score, OBD_i is an indicator for whether student i has observed background disadvantage, $\overline{OBD}_{jt}$ is the fraction of students with observed background disadvantage in the classroom taught by teacher j in year t, and η_{ijt} is an error term that comprises teachers' time invariant value-added ϕ_j, an idiosyncratic time-varying classroom effect, ε_{jt}, and an idiosyncratic time-varying student effect, υ_{it}.

Another common specification used often in the literature corresponds to our basic estimating equation in gains which has the following form:

$$T_{ijt} - T_{it-1} = \beta_0 + \beta_1 OBD_i + \beta_2 \overline{OBD}_{jt} + \eta_{ijt} \qquad (8)$$

It is important to note, however, that if we consider the model in levels described in (7) and subtract the lagged test scores from both sides of the equation we get:

$$T_{ijt} - T_{it-1} = \alpha_0 + \alpha_2 OBD_i + \alpha_3 \overline{OBD_{jt}} + (1-\alpha_1) T_{it-1} + \eta_{ijt} \qquad (9)$$

As we can see in equation (9) models in levels and in gains would then be equivalent only under the assumption that the persistence parameter α_1 equals 1. Otherwise the remaining unexplained part of the effect of lagged test scores will be an omitted variable leading to omitted variable bias in the estimated parameters, as lagged test scores will be correlated with the independent variables.

In the case of *random* sorting of teachers, the omitted variable (lagged test score) will only be correlated with the individual disadvantage indicator OBD_i. However, in the case of *non-random* sorting of teachers, lagged test score will also be correlated with classroom composition $\overline{OBD_{jt}}$ and with teacher effectiveness. Therefore, we expect the gains model to lead to biased estimates of the coefficient on the classroom composition and of teacher effects, epecially when using teacher fixed effect methods.

The first model that we consider, *aggregated residuals,* is one in which student achievement test scores are regressed on lagged test score and a student Observed Background

Disadvantage (OBD) indicator, as well as classroom-level averages of OBD status, which are intended to distinguish classroom composition effects from teacher effects. The student-level residuals are then aggregated by teacher, with each teacher's value-added defined as her students' average residual performance (see, for example, Kane & Staiger, 2005, 2008).

In the aggregated residuals model, we estimate equation (7) or (8) by OLS and compute the estimated residuals ($\hat{\eta}_{\mathrm{ijt}}$). We then obtain Empirical Bayes teacher-effect estimates, $\widehat{\phi}_{\mathrm{j}}$, by applying a random effects estimation in the following equation:

$$\hat{\eta}_{\mathrm{ijt}} = \phi_j + \zeta_{\mathrm{ijt}} \tag{10}$$

Note that in aggregating the residuals to estimate teachers' effectiveness, we assume that teachers' time-invariant effects (ϕ_{j}) are not correlated with the rest of explanatory variables in the model. Any true variance in teacher effectiveness associated with classroom-average characteristics may be attributed to the classroom-average covariates, potentially leading to biased estimates of teacher value-added. This is a key potential drawback of this type of model for estimating the correlation between teacher value-added and student characteristics, though the severity of the bias under varying assumptions is an empirical question.

The second class of model is a *random effects or variance-components model*, which assumes that a portion of the unexplained variation in students' test scores is associated with teacher effectiveness. The random effects model differs from the aggregated residuals model in that the presence of the teacher random effect in the error term causes the estimator to draw more on the within-teacher relationship between test scores and covariates. In such a model, the teacher value added is estimated by the best linear unbiased predictions (BLUP) from the model, which reflect the imprecision in each teacher's estimate due to the number of students taught and the correlation with classroom covariates (McCaffrey et al., 2003). However, like the aggregated residual estimates, *teacher random effects estimates* of equation (7) and (8) are obtained under the additional assumption that the teacher time-invariant effects are not correlated with the rest of explanatory variables in the model. Random effects estimators give us teacher effects that are already shrunken, and no further Empirical Bayes adjustments are necessary. As both aggregated residuals methods and teacher random effects estimates require that explanatory variables are not correlated with the teacher time-invariant effects, they will provide biased estimates when we deviate from the random assignment of teachers into classrooms. It is an empirical question, however, how this bias in the estimated coefficients translates into bias in the estimation of our correlation parameter of interest, ρ .

The third class of models consists of *teacher fixed effects models,* which are estimated by allowing teacher time invariant effects (ϕ_j) to be correlated with student and classroom characteristics. These models are then robust to non-random sorting of teachers into classrooms and are then expected to provide unbiased estimates of the model parameters under non-random sorting of teachers into classrooms. However, with small samples and limited variability of classroom composition, estimates based on teacher fixed effect models can suffer from problems of high noise in the estimated parameters of both classroom composition and teacher effects. This is so because the information available to distinguish between the effects of classroom composition and teacher effectiveness will be limited. It is important to point out that this won't be a problem with *aggregated residuals* or *teacher random effect models* because estimates in this case are based on both between and within teacher variation in the data.

However, there is a possible source of bias in the fixed effects estimator if teachers have some of the same students in consecutive years. It is well known that panel data models that include both *student* fixed effects and a lagged dependent variable lead to biased estimates

(Nickell, 1981). Although our specifications use *teacher* fixed effects and not *student* fixed effects, a similar bias will occur in our levels specification if some students have the same teacher in consecutive years. For these "repeater" students, their lagged test score will be correlated with the true teacher effect for their current year, a part of which will be in the error term for the estimated model. This will lead to bias in all the estimated coefficients in the equation. In our limited classroom variability scenario in which students and teachers stay in the same school, one-sixth of the students on average have the same teacher in consecutive years thereby leading to bias in the levels models with teacher fixed effects. Note that this is not a problem in the gains model, because the lagged test score is not included on the right hand side of the model. It is an empirical question then to assess the tradeoff between bias and variance implied by these estimators in different scenarios and study how these translate into estimates of our correlation parameter of interest ρ .

For the teacher fixed effects models, we apply Empirical Bayes adjustments to our teacher effects estimates following Tate (2004). Let $\hat{\sigma}_j^2$ be the estimated variance of the teacher effect estimate ϕ_j and $\bar{\mu}$ and $\hat{\sigma}^2$ be the mean and variance, respectively, of the distribution of estimated teacher effects in the sample. We construct the Empirical Bayes estimate of each teacher's value-added as:

$$\phi_j^{EB} = \hat{\phi}_j \cdot \chi_j + \bar{\mu} \cdot (1 - \chi_j) \tag{11}$$

where χ_j is the reliability of $\hat{\phi}_j$, which equals $\hat{\sigma}^2/(\hat{\sigma}^2 + \hat{\sigma}_j^2)$.

Finally, in order to study the sensitivity of the results to the inclusion of classroom average characteristics, all models are also estimated without including the controls for the average proportion of OBD students.

6. Results

6.1 Baseline scenario: Accuracy of teacher effects estimates and the distribution of effective teaching under random teacher assignment

In Table 2 we present average estimates from our Monte Carlo simulations for the case in which teacher assignment to classrooms is random. We begin by focusing on how accurate the different models are at recovering true teacher effects under various assumptions about variability in classroom composition and student learning decay. We measure accuracy by the rank correlation between true and estimated teacher effects $(Corr(\phi_j, \widehat{\phi_j}))$. All model specifications represented in Table 2 control for student-level OBD status and the classroom proportion OBD; models in levels also control for students' lagged test scores.

We highlight three findings about the accuracy of estimated teacher effects in recovering true teacher contributions to learning. Our first finding is that, regardless of model specification, assumptions about variability in classroom composition and about year-to-year student learning decay, models in levels produce more accurate estimates of teacher effects than models in gains. For instance, when we assume that there is extreme variability in classroom composition and that learning persistence (β) is 0.8, the average rank correlation between true and estimated teacher effects is 0.93 for models in levels and is 0.91 for models in gains (Panel A). At the other extreme, with limited variability in classroom composition and student learning persistence of 0.2, the average rank correlation is around 0.8 for models in levels and around 0.68 for models in gains (Panel C). A similar pattern holds for all other intermediate cases and modeling choices. It is not clear whether we would find a similar pattern in a scenario of no decay (represented here as β=1), since gains models assume zero decay (Clotfelter, Ladd, & Vigdor, 2007). However, work by Guarino, Reckase, and Wooldridge (2013) demonstrates strong performance of levels relative to gains models even under conditions of perfect learning persistence. Moreover, it is also not clear that an assumption of zero learning decay is reasonable in real-world contexts.

Our second finding is that, for any given modeling choice and assumption about variability in classroom composition, the accuracy of estimated teacher effects diminishes with the amount of student learning decay. In other words, it is easier to accurately predict a teacher's true contribution to learning when learning gains are persistent over time than when they are not. For example, in the scenario of extreme year-to-year variability in classroom composition under an aggregated residuals model, the rank correlation between true and estimated teacher-effects is 0.93 when the persistence parameter (representing 1 minus the rate of decay) is 0.8, is 0.88 when the persistence parameter is 0.4, and is 0.85 when the persistence parameter is 0.2 (column 1, Panel A of Table 2). Similarly, under limited variability in classroom composition, an empirical Bayes teacher fixed effects estimator in test-score gains produces a rank correlation with true teacher effects that is 0.85 with persistence of 0.8, is 0.74 with persistence of 0.4 and is 0.68 with persistence of 0.2 (column 8, Panel C of Table 2). The same pattern holds for other modeling choices in both levels and gains.

Our third finding concerning accuracy of teacher effect estimates under a baseline scenario of random teacher assignment is that, for all modeling choices and assumptions about student learning decay, estimates from settings with extreme variability in classroom composition are more accurate at recovering true teacher effects than those from settings with limited variability in classroom composition. This is specially the case for models using teacher fixed effects. As discussed above this is not surprising as with limited variability in classroom composition we are limiting the available within variation in the data that fixed effect methods are based on.

Continuing with the baseline case of random teacher assignment, we now turn attention to estimates of the distribution of effective teaching with respect to student background disadvantage. Here we highlight four main findings.

The first finding is that, regardless of the assumptions we consider about student learning decay, under extreme variability in classroom composition all models produce accurate estimates of the correlation between teacher effectiveness and the proportion of OBD students a teacher faces. In the case of random teacher assignment, this correlation is close to zero by construction, and we find that, on average, estimates of the distribution correlation are centered on zero.

The second finding is that, even when teacher assignment is random, estimates of the distribution correlation from teacher fixed effects models in levels (with and without post-hoc empirical Bayes adjustment) are negatively biased when there is limited variability in classroom composition regardless of student learning decay assumptions. With student learning persistence of 0.8, for instance, the estimated average distribution correlation is -0.24 without empirical-Bayes adjustment and -0.23 with adjustment (columns 3 and 4, Panel A of Table 2). At the other extreme, with persistence of 0.2, the estimated average distribution correlation is -0.25 with and without adjustment. We find some variation in the average estimates of the distribution correlation parameter depending on the degree of learning decay assumed, but such relation does not seem to be monotonic.

The third finding is that when teacher assignment is random and there is limited variability in classroom composition, aggregated residuals models in levels produce the most accurate estimates, on average, of the distribution correlation. Random-effects models in levels represent an intermediate case between aggregated residuals and teacher fixed effects, with estimates between -.08 and -0.14 depending on assumptions about student learning decay.

Our fourth finding is that when teacher assignment is random and there is limited variability is classroom composition, teacher random- and fixed effects models estimated in gains produce distribution correlations that are more accurate than those based on models in levels regardless of the assumptions about student learning decay. With learning persistence of 0.2, for example, the average distribution from random-effects models changes from -0.08 to 0.002 going from levels to gains estimation (columns 2 and 6, Panel C of Table 2). Similarly, for teacher fixed effects models with and without empirical Bayes adjustment, the average distribution correlation moves from -0.25 when we estimate the value-added model in levels to 0.014 when we estimate it in gains (columns 3, 4, 7 and 8, Panel C of Table 2).

One possible explanation for the second finding regarding the bias in teacher fixed effects models applied to samples with limited variability in classroom composition is that teacher effect estimates from models in levels are imprecise when there is limited variability of classroom composition. Our simulation results, however, do not support this idea. For any given level of student learning decay, the rank correlation between true and estimated teacher effects is always *weaker* in models in gains than in models in levels.

Another possible explanation is bias in estimates of the coefficients on the observed covariates of the value-added model. As Appendix Table A1 shows, however, this conjecture is also unlikely because, for models estimated in levels in settings with limited classroom variability, the bias in estimates of student learning persistence (β) and in estimates of the coefficient on individual background disadvantage status is similar for aggregated residuals, teacher random effects and teacher fixed effects models. Moreover, the bias in the coefficient estimate for the classroom proportion of disadvantaged students in teacher fixed effects models is similarly biased when we estimate the model in levels and in gains.

A plausible explanation for the results presented above stems from the fact that when we estimate teacher fixed effects models in settings with limited variability in classroom composition, even if teacher assignment is random, the covariance between contemporaneous classroom proportion of student disadvantage and lagged classroom average scores is non-zero. In large samples, with extreme variability in classroom composition, this covariance approaches zero, which is why all models in levels recover without bias the distribution correlation. With limited variability in classroom composition, a teacher gets assigned to a similar "type" of student year-after-year so that there is systematic correlation between classroom composition, teacher effectiveness and student learning over time. Limited variability in classroom composition is, therefore, not a concern in aggregated residuals models because by assuming that teacher effects are the portion of test scores that are uncorrelated with covariates included in the model, aggregated residuals is effectively imposing orthogonality between teacher value-added and classroom composition even though in the data-generating process, classroom composition contributes to learning. This problem would be potentially less important in models in gains, especially with lower levels of learning decay (β close to 1), because lagged scores are not part of the set of explanatory variables.

Another likely explanation for the bias of teacher fixed effects estimates of levels models in the case of limited variability of classroom composition comes from Nickell bias as discussed above. This bias will only arise when the lagged test score is on the right hand side, as is the case in the levels model, and when some students have the same teacher in two consecutive years, as is the case in the limited variability of classroom composition scenario. It should be noted that the bias also shows up in the random effects model but at a lower level, which can be understood by recalling that random effects can be constructed as a weighted average of fixed effects and pooled estimators.

6.2 Sensitivity to the exclusion of classroom composition controls under random teacher assignment

Although teacher effects from random- and fixed-effects models in levels are fairly accurate estimates of the true teacher contributions to student learning, they produce inaccurate estimates of the distribution correlation in settings with limited variability in classroom composition as it might become difficult to separate classroom composition effects from teacher effectiveness. In this subsection we study how results change if we omit controls for classroom composition in our analysis. Results for this case are presented in Table 3.

Our three key findings concerning the accuracy of teacher effect estimates also hold for models that do not control for classroom composition. Specifically, as results in Table 3 indicate: i) models in levels produce more accurate estimates of teacher effects than models in gains; ii) the accuracy of estimated teacher effects diminishes with the amount of student learning decay and iii) teacher effect estimates from settings with extreme variability in classroom composition are more accurate at recovering true teacher effects than those from settings with limited variability in classroom composition.

While not controlling for classroom composition does not affect our conclusions about the accuracy of teacher effect estimates under random teacher assignment and the various scenarios we consider, not doing so substantially affects some of our conclusions about the distribution of effective teachers with respect to student disadvantage. We highlight three main changes.

First, regardless of modeling choice and assumptions about student learning decay, not controlling for classroom composition leads to negatively biased estimates of the distribution correlation even in the case of extreme variability in classroom composition.

Second, as Table 3 shows, not controlling for classroom composition introduces considerable bias in estimates of the distribution correlation parameter ρ derived from aggregated residuals and teacher random effects methods when these models are estimated in levels, in the case of limited variability in classroom composition. Before, estimates of ρ from aggregated residuals models were centered around zero, whereas now the average estimate is $\widehat{\rho} = -0.18$. Without class composition controls, the bias in the distribution correlation from teacher random effects is as severe as in teacher fixed effects models. Moreover, for both of these methods, the bias in the distribution correlation becomes more severe with greater student learning decay.

This result is probably not surprising if we take into account that when classroom composition is part of the data-generating process but we omit it from our estimation equation. The direct contribution of classroom composition to student learning—which previously was netted out—now loads onto the estimated teacher effects, creating an spurious correlation with a teacher's proportion of disadvantaged students. Since in our simulations teacher quality positively affects learning while the classroom proportion of disadvantaged students negatively does so, the direction of the bias becomes negative.

Third, while estimates of the distribution correlation from value-added models in gains are considerably better than those from models in levels when there is limited variability in classroom composition, the improvement in these estimates is less than in the case in which we control for classroom composition. This result is true regardless of the assumptions we make about student learning decay.

6.3 Accuracy of teacher effects estimates and the distribution of effective teaching under partially systematic teacher assignment

We now present simulation results from a perhaps more realistic scenario in which teachers are assigned to classrooms in a way that teacher quality becomes correlated with the proportion of disadvantaged students a teacher has. In particular, we assume that the true correlation between teacher effectiveness and the proportion of OBD students is -0.5, such that the best teachers tend to be assigned to the classrooms with the lowest proportion of OBD students.

Introducing systematic teacher assignment does not fundamentally alter our three main conclusions about the accuracy of estimated teacher effects from those in the case of random teacher assignment. We continue to find that models in levels produce more accurate estimates of teacher effects than models in gains; that the accuracy of estimated teacher effects diminishes with the amount of student learning decay and that teacher effect estimates from settings with extreme variability in classroom composition are more accurate at recovering true teacher effects than those from settings with limited variability in classroom composition.

However, systematic teacher assignment fundamentally alters our conclusions about the appropriateness of various value-added models to recover the distribution of teacher effectiveness with respect to student background disadvantage. For example, as Table 4 indicates, in the case of extreme variability in classroom composition, all models—particularly aggregate residuals—in both levels and gains understate the degree of teacher sorting with respect to student disadvantage. This conclusion holds regardless of assumptions about student learning decay.

Second, unlike the case of random teacher assignment, when there is limited variability in classroom composition, teacher fixed effects models in levels—with and without empirical Bayes adjustment—produce the best estimates of the distribution correlation, followed by random effects and aggregated residuals. When year-to-year learning decay is low ($\beta = 0.8$) fixed effects models do a better job at recovering the true distribution correlation than when learning decay is high ($\beta = 0.2$) but even in the case of high learning decay, teacher fixed effects models in levels perform notably better than aggregated residuals and teacher random effects when there is limited variability in classroom composition.

As Appendix Table A2 shows, part of the reason why aggregated residuals models in levels perform so poorly at recovering the distribution correlation is that, when there is partially systematic teacher assignment and limited variability in classroom composition, aggregated residuals models overstate by orders of magnitude the true contribution of classroom composition to student learning. On average, the estimate of the coefficient on classroom proportion OBD is -0.96 when the true parameter is -0.3. In doing so, the aggregated residuals model is purging some of the true contribution of teacher effectiveness to learning, and because teacher effectiveness positively affects student learning, the resulting estimate of the distribution correlation is biased towards zero. Although this coefficient was also biased in the case of random assignment, when using aggregated residuals methods, the bias was not as severe as in this case.

Third, unlike the case of random teacher assignment, when there is limited variability in classroom composition, models in gains produce even worse estimates of the distribution correlation regardless of assumptions about student learning decay. The most dramatic departure from the true distribution correlation when estimating models in gains occurs for fixed effects models with and without empirical Bayes adjustment. As Appendix Table A2 shows, this result is not driven by the fact that teacher fixed effects models in levels produce less biased estimates of the coefficient on classroom proportion of disadvantaged students than models in gains. In fact, with limited variability in classroom composition, the bias in the estimates of the coefficient on classroom proportion of disadvantaged students is roughly the same in levels and gains, independent of assumptions about student learning decay.

6.4 Sensitivity to the exclusion of classroom composition controls with partially systematic teacher assignment

In Table 5 we present results on the accuracy of teacher effect estimates and of the distribution correlation parameter ρ when there is systematic sorting of teachers into classrooms based on the proportion of disadvantaged students, and we exclude controls for classroom composition. Results on teacher effects accuracy and ρ for the case of extreme variability in classroom composition are identical to those in Table 4 in which we controlled for the classroom proportion OBD. The main difference, however, arises in the case of limited variability in classroom composition for the aggregated residuals model.

As Table 5 indicates, when there is limited variability in classroom composition and we do not control for the classroom proportion of disadvantaged students, the accuracy of teacher effect estimates from the aggregated residuals model in both gains and levels increases substantially and now matches that of teacher random and fixed effects models. This improvement in accuracy is most pronounced when there is high student learning decay, as in Panel A of Table 5, than when we assume a rate of persistence of 0.8.

Not controlling for classroom composition also improves our estimates of ρ using aggregated residuals in cases with limited variability in classroom composition. For example, when persistence is 0.8 (Panel A) the average estimated ρ is now -0.39 which is much closer to the true value of -0.5 (when controlling for classroom composition our estimated ρ was only -0.06). This improvement in estimating ρ is even larger when student learning decays more. As Panels B and C of Table 5 indicate, the average estimated ρ is-0.46 when we set persistence to be 0.4 or 0.2 (column 1). As column 5 shows, not controlling for classroom composition also improves our estimated ρ when we estimate aggregated residuals models in gains, although our estimated ρ from models in gains continue to be significantly far from the true value when there is partially systematic teacher assignment and limited variability in classroom composition.

Not controlling for classroom composition also improves our estimates of ρ using teacher random effects methods value added models. As column 2 of Table 5 indicates, estimates of ρ from teacher random effects methods resemble those of teacher fixed effects models in this case. The improvement is most pronounced when there is little student learning decay, as is the case in Panel A of Table 5 when we assume persistence to be 0.8. However, like all other models, significant bias remains in estimates of ρ from teacher random effects in gains, regardless of the degree of student learning decay we assume.

7 Conclusion

In light of our finding from one large, urban district that the estimated distribution of teacher effectiveness among more and less disadvantaged students depended on our modeling approach, this simulation study sought to compare a variety of popular value-added models in terms of their ability to provide unbiased distribution estimates. Our parameters of interest were not only the rank correlation between true and estimated teacher value-added parameters, but also the correlation between estimated teacher value-added and the proportion of the teacher's students who are observably disadvantaged. Our simulations maintained the assumption that students' test scores were affected by their own disadvantage as well as by the average disadvantage of their classmates. Our simulations varied the amount of systematic teacher sorting, the amount of student learning decay, and the amount of within-teacher variation in classroom composition. Our estimates compare the accuracy of aggregated residual, teacher random effects, and teacher fixed effects models estimated in levels and in gains, with and without the inclusion of classroom composition covariates.

We find that the distribution of teacher value-added by student characteristics depends on modeling method and specification. Moreover, we find that models that do well at recovering the relative size of teachers' contributions to learning are not necessarily those that do a good job on estimating the degree of teacher sorting by student characteristics. Specifically, models in levels that employ teacher fixed effects produce less-biased estimates of student sorting than the other approaches when sorting is indeed nonrandom. However, in the absence of student sorting, aggregated residuals methods or models in gains perform better at recovering the distribution parameter. Since the true amount of sorting is unknown in real-world contexts, it may be preferable to exclude classroom characteristics from aggregated residuals or teacher random effects models, especially if demographic sorting is likely.

If using teacher fixed effects models, which appear to have particular advantages when sorting is present, the use of gains rather than levels appears preferable if the distribution correlation is the parameter of interest. This is particularly true when some or all students have the same teacher in consecutive years, as is the case with "looping" or with a single teacher teaching the advanced versions of a subject in multiple grades. But when the parameters of interest are the teacher fixed effects themselves, estimates from levels appear preferable in scenarios of low as well as high decay.

Given the sensitivity of both the rank correlations and the distribution parameter to the amount of true student sorting and to the within-teacher variability in classroom composition, the choice of model should potentially take contextual information into account. For instance, an analysis of variance of student characteristics between and within teachers will shed light on the amount of variability available from which to distinguish classroom and teacher effects. The amount of likely sorting is also an important consideration. Though schools may claim that students are randomly assigned to teachers and classrooms, it may be worth understanding the assignment procedures more carefully in a particular context to anticipate the amount of true sorting that may be present. If high sorting is likely, aggregated residuals may not be the best approach for recovering the distribution parameter. If high sorting is unlikely, however, teacher fixed effects may be less desirable, especially estimated in levels.

The results we present here simply underscore the importance of testing the results of any modeling choice for sensitivity to that choice and of transparently reporting that information to policymakers, alongside clear guidance about which models have the best properties in a particular context.

References

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

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics, 3*, 29-54.

Becker, H. S. (1952). The career of the Chicago public school teacher. *American Journal of Sociology, 57*(5), 470-477.

Borjas, G. J., & Sueyoshi, G. T. (1994). A two-stage estimator for probit models with structural group effects. *Journal of Econometrics*, *64*(1-2), 165-182.

Buddin, R. & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. Journal of Urban Economics, 66, 103-115.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (Working Paper # 17699). Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review, 24*(4), 377-392.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). How and why do teacher credentials matter for student achievement? (Working Paper # 12828). Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*, *45*(3), 655-681.

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. *Education Finance and Policy, 5*(3), 278-316. doi: 10.1162/EDFP_a_00002

Guarino, C., Reckase, M. D., & Wooldridge, J. M. (2013). Can value-added measures of teacher performance be trusted? Bloomington, IN: Indiana University.

Hannaway, J., Xu, Z., Sass, T. R., Figlio, D., & Feng, L. (2010, October 19). *Value added of teachers in high-poverty schools and lower poverty schools: Implications for research, policy and management.* Paper presented at the Association of Public Policy Analysis and Management Fall Conference, Boston, MA.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354.

Jacob, B. A., & Lefgren, L. (2007). Principals as agents: Subjective performance assessment in education. *Journal of Labor Economics, 26*(1), 101-136.

Kane, T. J., & Staiger, D. O. (2005). Using imperfect information to identify effective teachers (pp. 61). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Working paper no. 14607. Cambridge, MA: National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the education production function. Working paper. San Diego, CA: University of California San Diego.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability (pp. 191). Santa Monica, CA: RAND Corporation.

Nickell, S. (1981) Biases in Dynamic Models with Fixed Effects. *Econometrica* , 49(6), 1417-1426.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.

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

Tate, R. (2004). A cautionary note on shrinkage estimates of school and teacher effects. *Florida Journal of Educational Research, 42,* 1-21.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*(485), F3-F33.

Table 1. Empirical estimates of the distribution correlation between estimated teacher effectiveness and the proportion OBD of a teacher's students ($\hat{\rho}$) using data from a large urban school district.

	Levels			Gains		
	AR	RE	FE	AR	RE	FE
	(1)	(2)	(3)	(4)	(5)	(6)
With classroom controls	-0.064	-0.252	-0.227	0.007	0.015	-0.004
Without classroom controls	-0.133	-0.246	-0.207	0.029	0.04	0.014

Notes: All test scores are standardized within grade and year prior to estimation. AR=aggregated residuals model; RE=teacher random effects model; FE=teacher fixed effects estimation model

Table 2. Monte Carlo simulation estimates of the rank correlation between true and estimated teacher effects ($corr(\phi_j, \hat{\phi}_j)$) and of the distribution correlation between teacher effectiveness and proportion of OBD students ($\hat{\rho}$) under random teacher assignment (i.e. ρ=0) controlling for classroom average OBD.

		Levels				Gains			
		AR	RE	FE	EB-FE	AR	RE	FE	EB-FE
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variability of Classroom composition		**Panel A. Persistence=0.8**							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.925	0.926	0.926	0.926	0.911	0.911	0.911	0.911
	$\hat{\rho}$	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.916	0.906	0.882	0.880	0.862	0.860	0.852	0.848
	$\hat{\rho}$	-0.020	-0.144	-0.237	-0.230	-0.006	-0.036	-0.072	-0.070
		Panel B. Persistence=0.4							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.882	0.883	0.882	0.882	0.821	0.821	0.820	0.821
	$\hat{\rho}$	-0.001	-0.001	-0.002	-0.002	0.001	0.001	0.000	0.000
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.870	0.865	0.824	0.823	0.758	0.758	0.746	0.744

	$\hat{\rho}$	-0.028	-0.112	-0.285	-0.277	-0.004	-0.010	-0.041	-0.040
		Panel C. Persistence=0.2							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.850	0.850	0.850	0.850	0.768	0.768	0.768	0.767
	$\hat{\rho}$	0.003	0.003	0.003	0.003	0.007	0.009	0.010	0.010
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.835	0.832	0.794	0.793	0.697	0.697	0.679	0.677
	$\hat{\rho}$	-0.024	-0.078	-0.254	-0.246	0.001	0.002	0.014	0.014

Notes: This table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals model; RE= teacher random effects model; FE=teacher fixed effects model; EB-FE=empirical Bayes teacher fixed effects model. In the extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In the limited variability of classroom composition scenario we assume that there are 20 schools, no student or teacher mobility across schools and teachers and students are reassigned randomly to classrooms each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in the data generating process.

Table 3. Monte Carlo simulation estimates of the rank correlation between true and estimated teacher effects ($corr(\phi_j, \hat{\phi}_j)$) and of the distribution correlation between teacher effectiveness and proportion of OBD students ($\hat{\rho}$) under random teacher assignment (i.e. ρ=0) without controlling for classroom average OBD.

		Levels				Gains			
		AR	RE	FE	EB-FE	AR	RE	FE	EB-FE
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variability of Classroom composition		**Panel A. Persistence=0.8**							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.922	0.922	0.922	0.922	0.907	0.907	0.907	0.907
	$\hat{\rho}$	-0.074	-0.075	-0.075	-0.075	-0.072	-0.073	-0.073	-0.073
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.899	0.884	0.882	0.882	0.859	0.857	0.857	0.856
	$\hat{\rho}$	-0.183	-0.251	-0.256	-0.255	-0.069	-0.093	-0.096	-0.096
		Panel B. Persistence=0.4							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.88	0.88	0.879	0.88	0.819	0.819	0.819	0.819
	$\hat{\rho}$	-0.07	-0.071	-0.071	-0.071	-0.066	-0.067	-0.067	-0.067
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.848	0.832	0.827	0.827	0.757	0.757	0.756	0.757

	$\hat{\rho}$	-0.221	-0.282	-0.297	-0.296	-0.033	-0.041	-0.045	-0.044
		Panel C. Persistence=0.2							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.849	0.848	0.848	0.849	0.767	0.767	0.767	0.767
	$\hat{\rho}$	-0.064	-0.065	-0.065	-0.065	-0.057	-0.058	-0.058	-0.059
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.815	0.802	0.796	0.796	0.697	0.696	0.695	0.696
	$\hat{\rho}$	-0.212	-0.265	-0.284	-0.283	-0.014	-0.017	-0.019	-0.019

Notes: This table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals model; RE= teacher random effects model; FE=teacher fixed effects model; EB-FE=empirical Bayes teacher fixed effects model. In the extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In the limited variability of classroom composition scenario we assume that there are 20 schools, no student or teacher mobility across schools and teachers and students are reassigned randomly to classrooms each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in the data generating process.

Table 4. Monte Carlo simulation estimates of the rank correlation between true and estimated teacher effects ($corr(\phi_j, \hat{\phi}_j)$) and of the distribution correlation between teacher effectiveness and proportion of OBD students ($\hat{\rho}$) under partially systematic teacher assignment of the most effective teachers to low-proportion OBD classrooms (i.e. ρ=-0.5) controlling for classroom average OBD.

		Levels				Gains			
		AR	RE	FE	EB-FE	AR	RE	FE	EB-FE
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variability of Classroom composition		**Panel A. Persistence=0.8**							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.915	0.925	0.926	0.925	0.898	0.908	0.909	0.909
	$\hat{\rho}$	-0.261	-0.357	-0.366	-0.364	-0.255	-0.345	-0.359	-0.356
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.846	0.912	0.912	0.910	0.794	0.819	0.842	0.838
	$\hat{\rho}$	-0.058	-0.321	-0.506	-0.496	-0.030	-0.105	-0.210	-0.205
		Panel B. Persistence=0.4							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.867	0.877	0.878	0.877	0.808	0.817	0.820	0.819
	$\hat{\rho}$	-0.253	-0.335	-0.358	-0.356	-0.234	-0.297	-0.334	-0.332
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.813	0.857	0.850	0.850	0.689	0.695	0.716	0.715

	$\hat{\rho}$	-0.071	-0.246	-0.595	-0.580	-0.025	-0.043	-0.153	-0.149
		Panel C. Persistence=0.2							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.833	0.841	0.843	0.843	0.749	0.756	0.760	0.760
	$\hat{\rho}$	-0.246	-0.314	-0.346	-0.343	-0.219	-0.266	-0.311	-0.309
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.778	0.812	0.831	0.831	0.624	0.625	0.624	0.623
	$\hat{\rho}$	-0.069	-0.189	-0.572	-0.561	-0.016	-0.021	-0.063	-0.061

Notes: This table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals model; RE= teacher random effects model; FE=teacher fixed effects model; EB-FE=empirical Bayes teacher fixed effects model. In the extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In the limited variability of classroom composition scenario we assume that there are 20 schools, no student or teacher mobility across schools and teachers and students are reassigned randomly to classrooms each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in the data generating process.

Table 5. Monte Carlo simulation estimates of the rank correlation between true and estimated teacher effects ($corr(\phi_j, \hat{\varphi}_j)$) and of the distribution correlation between teacher effectiveness and proportion of OBD students ($\hat{\rho}$) under partially systematic teacher assignment of the most effective teachers to low-proportion OBD classrooms (i.e. ρ=-0.5) without controlling for classroom average OBD.

		Levels				Gains			
		AR	RE	FE	EB-FE	AR	RE	FE	EB-FE
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variability of Classroom composition		**Panel A. Persistence=0.8**							
Extreme	$corr(\phi_j, \hat{\varphi}_j)$	0.926	0.926	0.926	0.926	0.909	0.909	0.909	0.909
	$\hat{\rho}$	-0.364	-0.368	-0.369	-0.367	-0.356	-0.36	-0.361	-0.359
Limited	$corr(\phi_j, \hat{\varphi}_j)$	0.920	0.917	0.916	0.915	0.840	0.854	0.856	0.855
	$\hat{\rho}$	-0.386	-0.511	-0.521	-0.52	-0.173	-0.225	-0.232	-0.231
		Panel B. Persistence=0.4							
Extreme	$corr(\phi_j, \hat{\varphi}_j)$	0.878	0.878	0.878	0.878	0.82	0.82	0.82	0.82
	$\hat{\rho}$	-0.353	-0.357	-0.358	-0.357	-0.328	-0.331	-0.333	-0.332

Limited	$corr(\phi_j, \hat{\phi}_j)$	0.877	0.863	0.856	0.856	0.71	0.716	0.719	0.719
	$\hat{\rho}$	-0.463	-0.571	-0.597	-0.595	-0.09	-0.107	-0.12	-0.119
		Panel C. Persistence=0.2							
Extreme	$corr(\phi_j, \hat{\phi}_j)$	0.843	0.844	0.844	0.844	0.760	0.760	0.761	0.761
	$\hat{\rho}$	-0.340	-0.343	-0.345	-0.343	-0.304	-0.307	-0.31	-0.308
Limited	$corr(\phi_j, \hat{\phi}_j)$	0.849	0.842	0.836	0.835	0.634	0.637	0.639	0.639
	$\hat{\rho}$	-0.461	-0.553	-0.587	-0.586	-0.048	-0.055	-0.062	-0.062

Notes: This table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals model; RE= teacher random effects model; FE=teacher fixed effects model; EB-FE=empirical Bayes teacher fixed effects model. In the extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In the limited variability of classroom composition scenario we assume that there are 20 schools, no student or teacher mobility across schools and teachers and students are reassigned randomly to classrooms each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in the data generating process.

Appendix Table A1. Monte Carlo simulation value added model parameter estimates under random teacher assignment (i.e. ρ=0) controlling for classroom average OBD. The true value for the coefficient on OBD status (α) is -0.08; for the coefficient on class proportion OBD (λ) is -0.3.

		Levels			Gains		
		AR	RE	FE	AR	RE	FE
Variability of Classroom composition				**Panel A. Persistence (β) =0.8**			
Extreme	$\hat{\beta}$	0.329	0.328	0.328	-	-	-
	$\hat{\alpha}$	-0.137	-0.138	-0.138	-0.027	-0.027	-0.027
	$\hat{\lambda}$	-0.717	-0.719	-0.719	-0.712	-0.712	-0.713
					-	-	-
Limited	$\hat{\beta}$	0.364	0.341	0.339			
	$\hat{\alpha}$	-0.141	-0.145	-0.145	-0.055	-0.055	-0.055
	$\hat{\lambda}$	-0.459	-0.221	-0.028	-0.167	-0.108	-0.037
				Panel B. Persistence (β) =0.4			
Extreme	$\hat{\beta}$	0.023	0.023	0.023	-	-	-
	$\hat{\alpha}$	-0.119	-0.119	-0.119	0.004	0.004	0.004
	$\hat{\lambda}$	-0.452	-0.449	-0.448	-0.466	-0.464	-0.463
Limited	$\hat{\beta}$	0.037	0.025	0.023	-	-	-
	$\hat{\alpha}$	-0.117	-0.118	-0.118	-0.017	-0.017	-0.017

	$\hat{\lambda}$	-0.378	-0.265	-0.007	-0.054	-0.046	-0.003
		Panel C. Persistence (β) =0.2					
Extreme	$\hat{\beta}$	0.003	0.003	0.003	-	-	-
	$\hat{\alpha}$	-0.098	-0.098	-0.098	0.013	0.013	0.013
	$\hat{\lambda}$	-0.365	-0.369	-0.369	-0.386	-0.399	-0.407
Limited	$\hat{\beta}$	0.011	0.003	0.000	-	-	-
	$\hat{\alpha}$	-0.091	-0.092	-0.092	-0.003	-0.003	-0.003
	$\hat{\lambda}$	-0.307	-0.247	-0.031	-0.025	-0.025	-0.039

Notes: This table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals model; RE= teacher random effects model; FE=teacher fixed effects model; EB-FE=empirical Bayes teacher fixed effects model. In the extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In the limited variability of classroom composition scenario we assume that there are 20 schools, no student or teacher mobility across schools and teachers and students are reassigned randomly to classrooms each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in the data generating process.

Appendix Table A2. Monte Carlo simulation value added model parameter estimates under partially systematic teacher assignment of the most effective teachers to low-proportion OBD classrooms (i.e. ρ=-0.5) controlling for classroom average OBD. The true value for the coefficient on OBD status (α) is -0.08; for the coefficient on class proportion OBD (λ) is -0.3.

		Levels			Gains		
		AR	RE	FE	AR	RE	FE
Variability of Classroom composition		**Panel A. Persistence (β) =0.8**					
Extreme	$\hat{\beta}$	0.329	0.327	0.327	-	-	-
	$\hat{\alpha}$	-0.132	-0.132	-0.132	-0.015	-0.015	-0.015
	$\hat{\lambda}$	-0.951	-0.106	-0.021	-0.941	-0.150	-0.020
Limited	$\hat{\beta}$	0.378	0.357	0.353	-	-	-
	$\hat{\alpha}$	-0.140	-0.143	-0.143	-0.050	-0.050	-0.050
	$\hat{\lambda}$	-0.955	-0.467	-0.021	-0.360	-0.225	-0.031
		Panel B. Persistence (β) =0.4					
Extreme	$\hat{\beta}$	0.024	0.024	0.024	-	-	-
	$\hat{\alpha}$	-0.119	-0.119	-0.119	0.011	0.011	0.011
	$\hat{\lambda}$	-0.616	-0.141	0.001	-0.611	-0.229	0.003
Limited	$\hat{\beta}$	0.040	0.027	0.022	-	-	-
	$\hat{\alpha}$	-0.115	-0.116	-0.116	-0.010	-0.010	-0.010
	$\hat{\lambda}$	-0.814	-0.593	0.020	-0.114	-0.091	0.052

		Panel C. Persistence (β) =0.2					
Extreme	$\hat{\beta}$	0.004	0.004	0.004	-	-	-
	$\hat{\alpha}$	-0.101	-0.101	-0.101	0.014	0.014	0.014
	$\hat{\lambda}$	-0.480	-0.151	0.006	-0.470	-0.226	0.012
Limited	$\hat{\beta}$	0.015	0.006	0.001	-	-	-
	$\hat{\alpha}$	-0.094	-0.095	-0.096	-0.006	-0.006	-0.006
	$\hat{\lambda}$	-0.677	-0.551	-0.007	-0.048	-0.043	0.006

Notes: Table shows average estimates from 100 Monte Carlo simulations with four years of student test score data, 120 teachers a single cohort of 2,400 students and data generating process following the parameterization described in the text. AR=aggregated residuals estimation model; RE=random effects estimation model; FE=fixed effects estimation model. In extreme variability of classroom composition scenario we assume that there is only one school and all teachers and students are randomly reassigned to classrooms each year. In limited variability of classroom composition scenario we assume that there are twenty schools, no student mobility across schools and teachers and students are reassigned randomly each year within school. Various persistence scenarios refer to the assumptions about the true coefficient of the lagged score in data generating process.