NBER WORKING PAPER SERIES

USING LAGGED OUTCOMES TO EVALUATE BIAS IN VALUE-ADDED MODELS

Raj Chetty
John N. Friedman
Jonah Rockoff

Working Paper 21961
http://www.nber.org/papers/w21961

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2016

We thank Gary Chamberlain, Guido Imbens, Patrick Kline, and Jesse Rothstein for comments. On May 4, 2012, Chetty was retained as an expert witness by Gibson, Dunn, and Crutcher LLP to testify about the importance of teacher effectiveness for student learning in Vergara v. California, a case that was decided on June 10, 2014, before research on this paper began. On June 13, 2015, Friedman was retained as a potential expert witness by the Radey Law Firm to advise the Florida Department of Education on the importance of teacher effectiveness for student learning, as related the development of Draft Rule 6A-5.0411. His work on this case lasted for approximately one week and he had no involvement in any legal proceeding. On December 11, 2015, Friedman was retained as an expert witness by the Houston Independent School District in the matter of Houston Federation of Teachers, et al. v. Houston Independent School District. Friedman has not shared or discussed this paper with the parties in that case or their attorneys. This research was funded in part by the National Science Foundation.

Using Lagged Outcomes to Evaluate Bias in Value-Added Models
Raj Chetty, John N. Friedman, and Jonah Rockoff
NBER Working Paper No. 21961
February 2016
JEL No. C18,H75,I21,J01,J08,J45,M50,M54

ABSTRACT

Value-added (VA) models measure the productivity of agents such as teachers or doctors based on the outcomes they produce. The utility of VA models for performance evaluation depends on the extent to which VA estimates are biased by selection, for instance by differences in the abilities of students assigned to teachers. One widely used approach for evaluating bias in VA is to test for balance in lagged values of the outcome, based on the intuition that today's inputs cannot influence yesterday's outcomes. We use Monte Carlo simulations to show that, unlike in conventional treatment effect analyses, tests for balance using lagged outcomes do not provide robust information about the degree of bias in value-added models for two reasons. First, the treatment itself (value-added) is estimated, rather than exogenously observed. As a result, correlated shocks to outcomes can induce correlations between current VA estimates and lagged outcomes that are sensitive to model specification. Second, in most VA applications, estimation error does not vanish asymptotically because sample sizes per teacher (or principal, manager, etc.) remain small, making balance tests sensitive to the specification of the error structure even in large datasets. We conclude that bias in VA models is better evaluated using techniques that are less sensitive to model specification, such as randomized experiments, rather than using lagged outcomes.

Raj Chetty
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
chetty@stanford.edu

Jonah Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

John N. Friedman
Department of Economis
Robinson Hall
Brown University
Providence, RI 02912
and NBER
john_friedman@brown.edu

Value-added (VA) models – which measure agents' productivity based on the outcomes they produce – are increasingly used to evaluate the performance of agents and institutions ranging from teachers and schools to doctors and CEOs. The utility of VA models for performance evaluation depends critically on the extent to which VA estimates are biased by selection, for instance by differences in the latent abilities of students assigned to teachers. Biased VA measures may systematically reward or penalize agents based on factors unrelated to true differences in performance.

One approach for evaluating bias in VA is to test for balance in lagged values of the outcome (e.g., Rothstein 2010). For example, many studies test whether students' prior scores are correlated with their current teachers' VA. This test is intuitive – today's inputs cannot influence yesterday's outcomes – and can be easily implemented in panel datasets typically used to estimate VA. Such tests also follow the precedent of examining pre-trends in outcomes to evaluate selection bias in the program evaluation literature.

We show that, despite their intuitive appeal, balance tests using lagged outcomes do not yield robust information about bias in value-added models. We use Monte Carlo simulations to demonstrate that lagged outcomes may be correlated with VA estimates even when VA estimates are unbiased. More generally, tests using lagged outcomes are uninformative about the degree of bias in misspecified VA models.

Tests of balance using lagged outcomes are less robust in VA applications than in conventional treatment effect settings (e.g., studies of variation in class size) for two reasons. First, the treatment itself (value-added) is estimated, rather than exogenously observed. As a result, correlated shocks to outcomes can enter both current VA estimates and lagged outcomes in non-transparent ways that depend on the error structure of the VA model. Second, in most VA applications, estimation error does not vanish in large datasets because the sample size per teacher (or principal, manager, etc.) remains small as the number of teachers grows large. Thus, the specification of the model's error structure remains important for inference even asymptotically. We conclude that bias in value-added models is better evaluated using techniques that are less sensitive to model specification, such as randomized experiments, rather than using lagged outcomes.

I. Model Setup

We consider the estimation of teachers' effects on students' test scores, but the results apply to other value-added applications. This section describes the data-generating process for student scores, the estimator for teacher VA, and the way we define bias.

Data-Generating Process. Let $c(i,t)$ denote student i's classroom in year t and $j(c(i,t)) = j(i,t)$ student i's teacher in year t. There are I students in each class. Each teacher teaches C classes per year, all in the same grade.[1] Each teacher has time-invariant value-added μ_j, drawn from a distribution with a mean normalized to 0.

We assess the sensitivity of tests for bias to model specification by introducing a "track" level shock that generates correlated errors in scores across grades. In particular, students and teachers are grouped into tracks (denoted by s). For instance, tracks might denote honors or remedial education tracks within a school. For simplicity, we assume that students and teachers are assigned to the same track in all years. There are common shocks to test scores across classrooms within the same track, both within and across grades. Such common shocks might arise from a good match between the honors curriculum and the material on statewide standardized tests in a particular year, or a school-wide initiative to improve remedial instruction.

Student's i's score in year t is

$$A_{it} = \delta_i + \alpha_i t + \mu_{j(i,t)} + \theta_{c(i,t),t} + \psi_{s(i),t} + \varepsilon_{it}, \tag{1}$$

where δ_i and α_i denote the idiosyncratic level and trend, respectively, of student i's scores. In addition to teacher VA μ_j, scores are affected by three random shocks: θ_{ct} at the classroom level, ψ_{st} at the track level, and ε_{it} at the individual level. These shocks are independently and identically distributed across classes, tracks, or individuals with mean 0 and variances σ^2_θ, σ^2_ψ, and σ^2_ε, respectively. Importantly, the teacher $j(i,t)$ to whom a student is assigned may depend upon the student's ability level δ_i and trend α_i. Following Rothstein (2010), we refer to sorting on δ_i as "static" sorting and sorting on α_i as "dynamic" sorting.

Value-Added Estimation. We use data on test scores and classroom assignments in two years $t \in \{1,2\}$ to forecast teacher quality in subsequent years. We estimate teacher VA using a gains

[1] For simplicity, we assume that all students progress through grades at a standard pace, so that we can omit notation for grade g (as it is collinear with year t for any given student).

specification.[2] Let $\Delta A_{it} = A_{it} - A_{i,t-1}$ denote student i's test score gain in year t and let $\Delta \bar{A}_{jt} = \sum_{i \in j} \Delta A_{it}$ denote the average gain for students taught by teacher j in year t. Our estimator for teacher j's VA is the average test score gain of the students they taught, scaled by a shrinkage factor to account for noise:

$$\hat{\mu}_j = \lambda \Delta \bar{A}_{j,t=2}, \tag{2}$$

where the shrinkage factor $\lambda = \sigma^2_{\mu} / \sigma^2_{\Delta \bar{A}_{jt}}$ represents the fraction of the variance in test score gains across teachers due to teacher effects. Following the approach of Kane and Staiger (2008), we estimate the variance components needed to calculate λ from the covariance of test score gains across classrooms (see the Appendix for details). The estimator in (2) is widely used because it minimizes the mean-squared error of out-of-sample forecasts of teacher quality (see e.g., Chetty, Friedman and Rockoff 2014).

Bias in VA Estimates. Suppose we randomly assign students to teachers in year $t = 3$ and regress their test score gains on their teachers' VA estimates based on observational data in years $t \in \{1,2\}$:

$$\Delta A_{it} = a + \beta \hat{\mu}_{j(i,t)} + \zeta_{it}. \tag{3}$$

The coefficient β in this regression identifies the degree of "forecast bias" in the VA model, defined as $b = 1 - \beta$ (Kane and Staiger 2008, Chetty, Friedman, and Rockoff 2014). If VA estimates are forecast unbiased ($b = 0$), assigning a student to a teacher with one unit higher value-added will, on average, increase her score by one unit.

The VA estimator in (2) yields forecast-unbiased estimates of teacher quality with static but not dynamic sorting of students to teachers. Sorting of students by ability levels is differenced out in test score gains, whereas trends in ability are not. The key issue is how to distinguish between these forms of sorting and, more generally, estimate the amount of forecast bias. In particular, do balance tests using lagged test score gains provide information about the degree of bias? We explore these questions in the next section.

[2] We use a gains specification rather than controlling for lagged test scores for simplicity; the qualitative results below naturally extend to VA estimators that control for lagged scores more flexibly.

II. Using Prior Scores to Evaluate Bias: Simulation Results

We simulate data on test scores using the data generating process described above with Normal distributions for each of the random variables in (1). The parameters we use in our baseline simulations are listed in Appendix Table 1 and the Stata code used for the simulations is available online. We estimate VA for each teacher using test score data from the first two years and study how these estimates predict test scores for a new cohort of students in year $t = 3$.

We begin with a baseline case where students are sorted to teachers on ability levels but not on trends, so the VA estimates in (2) are unbiased. When students are randomly assigned to teachers in year $t = 3$, estimating the regression specification in (3) yields a coefficient of $\beta = 1$, confirming that VA estimates are forecast unbiased (Table 1, Column 1).

Now consider using lagged outcomes to assess the degree of forecast bias when students are assigned to teachers in year $t = 3$ using the same (non-random) assignment process as in previous periods. We regress lagged gains $\Delta A_{i,t-1}$ on current teacher VA:

$$\Delta A_{i,t-1} = a + b' \hat{\mu}_{j(i,t)} + \nu_{it}. \tag{4}$$

Specifications of this form have been used to evaluate forecast bias in several studies, including our own prior work (Chetty, Friedman and Rockoff 2014, Bacher-Hicks, Kane and Staiger 2014, Rothstein 2015a). These tests are based on the idea that a non-zero coefficient b' in (4) constitutes evidence of bias since current teacher quality cannot have a causal effect on past test score growth. Contrary to this intuition, estimating this regression yields a coefficient of $b' = 0.7$ (Table 1, Column 2), even though VA estimates are unbiased in our simulation.

Why are lagged test score gains correlated with teacher VA even in the absence of bias? The reason is that the track level shock ψ_{st} enters both the VA estimate and the lagged gain, inducing a correlation between the two variables that is driven by noise rather than sorting. For example, consider estimating VA for 6th grade teachers in 1996 based on their students' test score gains in 1995. Suppose a positive track level shock in 1995 led to unusually high test scores in all grades in that year. This shock will artificially inflate VA estimates for 6th grade teachers. It will also increase 5th grade scores in 1995, driving up lagged score gains for 6th graders in 1996. As a result, VA estimates and lagged score gains will be positively correlated.

More generally, the coefficient on lagged scores in equation (4) is governed by the magnitude

of track level shocks:

$$b' = Cov\left(\Delta A_{i,t-1}, \hat{\mu}_{j(i,t)}\right) / Var(\hat{\mu}_{j(i,t)}) = \lambda Var\left(\psi_{st} - \psi_{s,t-1}\right) / Var(\hat{\mu}_{j(i,t)}) \geq 0. \quad (5)$$

In the special case without track shocks, where $Var\left(\psi_{st} - \psi_{s,t-1}\right) = 0$, the lagged score test would correctly diagnose the lack of bias ($b' = 0$). But when there are common shocks within tracks – or, more generally, any correlated errors in scores across grades – lagged gains will appear unbalanced across teachers with different VA estimates even when VA estimates are unbiased.

The fundamental reason that the lagged outcome balance test fails is that the treatment effect (teacher VA) is itself estimated using prior test score data.[3] If one observed each teacher's VA directly, the test would work as expected. When we use true VA $\mu_{j(i,t)}$ instead of $\hat{\mu}_{j(i,t)}$ when estimating (4), we obtain a coefficient of $b' = 0$ (Table 1, Column 3). This result shows that lagged outcomes are useful in testing for bias so long as one does not have to estimate the treatment effect itself, as in conventional treatment effect analyses (e.g., variation in class size).

Misspecified VA Models. We now turn to the case in which the model used to estimate VA is misspecified relative to the true data generating process for test scores. Suppose that students are sorted to teachers on trends (α_i) but that the econometrician continues to estimate VA using (2), ignoring dynamic sorting. We use simulations analogous to those above to explore whether the relationship between lagged gains and teacher VA is informative about the degree of bias in this setting.

In the absence of track shocks, regressing lagged gains on current teacher VA as in (4) yields a coefficient $b' = b$, the true degree of forecast bias. However, lagged gains are no longer informative about bias with track level shocks. In Figure 1, we plot b' (estimated as in Column 2 of Table 1) and b (estimated as in Column 1 of Table 1) vs. the degree of sorting on trends, measured by the correlation between μ_j and α_i. Estimates of forecast bias from the lagged gains regression (shown in triangles) not only differ in levels from the true values of forecast bias (in diamonds), but move in different directions as the degree of dynamic sorting changes.

[3] This estimation error persists even in large datasets because the number of observations per teacher typically does not grow with the sample size.

The preceding analysis assumes the econometrician takes track shocks into account when estimating VA, but ignores them when testing for balance in lagged outcomes. Naturally, one can modify the lagged outcome test to account for track level shocks. For example, subtracting the variance due to track level shocks yields an adjusted version of the lagged score regression coefficient,

$$b_{adj} = b' - \frac{Var(\psi_{st} - \psi_{s,t-1})}{Var(\hat{\mu}_{j(i,t)})}, \tag{6}$$

which identifies the true amount of forecast bias correctly ($b_{adj} = b$), as shown by the series in circles in Figure 1a. The problem with this approach is that it relies on specifying the model that generates test scores correctly. If the model for test scores is misspecified, the correction in (6) no longer works. Suppose, for instance, that track level shocks actually have a correlation of $\rho < 1$ across grades in practice, but the econometrician assumes that they are perfectly correlated across grades *both* when estimating VA and when implementing the lagged outcome test. Figure 1b replicates Figure 1a in this misspecified model. Here, even the adjusted estimate b_{adj} differs from true bias b because the correction in (6) is no longer valid.

Similar issues arise with other types of corrections, such as using one set of years to estimate VA and a distinct set of years to test for balance in prior scores. This approach yields a coefficient $b' = b$ under the model specified above because it ensures that the track level shocks that enter the VA estimate do not enter lagged gains. But once again, if the model is misspecified – for instance, because track shocks are serially correlated across years rather than iid – such an approach fails.

Of course, misspecification of the model for test scores will also generally lead to bias in VA estimates. For instance, failing to account for serially correlated track shocks will lead to forecast biased estimates of VA. The key point here is that tests for balance using lagged outcomes do not provide robust guidance on the true degree of bias in VA in such scenarios. Put differently, it is difficult to falsify any given hypothesis about the degree of forecast bias using data on lagged values of the outcome when one admits plausible forms of model misspecification.

Variants of the Lagged Outcome Test. One variant of the test above is to ask whether the forecast coefficient on value-added β changes when one controls for lagged gains when estimating the relationship between test score gains and teacher VA in observational data:

$$\Delta A_{it} = a + \beta \hat{\mu}_{j(i,t)} + \Delta A_{i,t-1} + \varepsilon_{it}. \tag{7}$$

Again, this test has intuitive appeal; if the inclusion of lagged gains $\Delta A_{i,t-1}$ affects β, this suggests that $\Delta A_{i,t-1}$ is an omitted variable correlated with $\hat{\mu}_{j(i,t)}$ that leads to bias in forecasting teacher's causal effects. We investigate the effect of controlling for lagged gains in Columns 4 and 5 of Table 1 using a set of students who are assigned to teachers in year $t = 4$ using the same (non-random) assignment process as in earlier periods.[4] In our baseline simulation with no dynamic sorting, a univariate regression of ΔA_{it} on $\hat{\mu}_{j(i,t)}$ yields a coefficient of $\beta = 1$ (Column 4). Controlling for $\Delta A_{i,t-1}$ reduces the coefficient to $\beta = 0.83$ even though VA estimates are actually unbiased. Mechanically, the problem is that the covariance between lagged gains and VA is driven by a component of lagged gains – the transitory track shock in year 2 ($\psi_{s,2}$) – that is uncorrelated with current gains. However, the OLS regression in (7) applies the cross-sectional correlation between lagged gains and current gains when partialling out the effect of $\Delta A_{i,t-1}$, biasing β away from 1. Once again, this problem arises because VA is estimated; if one uses true VA when estimating (7), controlling for $\Delta A_{i,t-1}$ does not affect β.

Another variant of the lagged outcome balance test is to examine whether there is excess variance in lagged gains across teachers.[5] Intuitively, we would not expect current teacher assignments to predict lagged gains in the absence of dynamic sorting. This test is typically implemented using an F-test in a regression of lagged gains on teacher fixed effects (Rothstein 2010). Such an F-test rejects the hypothesis of no teacher effects on lagged gains in our simulations even in the absence of dynamic sorting (Table 1, Column 2). The reason is that the standard F-test does not account for the correlation in lagged gains across classrooms taught by the same teacher due to track level shocks. Of course, one could implement a modified version of the test that accounts for these correlated errors, but this once again demonstrates the sensitivity of tests using lagged outcomes to model specification. Tests for excess variance are especially sensitive to the specification of the error structure in value-added applications because the number of students and classrooms per teacher typically does not grow with the sample size. Thus, the variance in test scores across

[4] We use $t = 4$ for this exercise so that current test score gains ΔA_{it} (which use data from $t = 3$ and $t = 4$) do not overlap with the data used to estimate value-added (which use data from $t = 1$ and $t = 2$).

[5] This test for "teacher-level bias" is more stringent than the tests for forecast bias discussed above because it seeks to determine whether VA measures are unbiased for every teacher rather than on average.

teachers is partly driven by noise even in large datasets.

III. Discussion

This paper has shown that tests of balance using lagged outcomes are sensitive to model specification in value-added applications because of estimation error in VA that persists asymptotically in large datasets. If one specifies the model for test scores correctly, one can formulate tests of balance using lagged outcomes that provide accurate measures of bias. However, test scores and other outcomes of interest are typically generated by a complex set of factors, making model misspecification quite likely. In such situations, tests of balance using lagged outcomes do not provide a reliable guide to the degree of bias in VA.

How can we estimate bias in a manner that is less sensitive to model specification? Conceptually, one needs data on test scores that are guaranteed to be uncorrelated with estimation error in VA irrespective of the underlying model for test scores. One way to obtain such a guarantee is to study experiments or quasi-experiments where students are assigned to teachers randomly (Kane and Staiger 2008, Cantrell and Kane 2013, Chetty, Friedman and Rockoff 2014). This approach can be expensive to implement and may generate imprecise estimates of bias due to limitations in power. Despite these challenges, several recent studies have used experimental and quasi-experimental methods to obtain informative estimates of forecast bias that turn out to be quite stable across settings, both for teacher value-added (see Glazerman and Protik 2015 for a survey) and school value-added (Bifulco, Cobb and Bell 2009, Deming 2014, Angrist et al. 2015).

Our exploratory analysis of model misspecification suggests several directions for future research. First, developing VA models that are robust to misspecification would be valuable, especially given the highly heterogeneous settings across which VA models are now being applied in education and other fields. Second, rather than testing the sharp null that VA estimates are unbiased – a knife-edge scenario that is unlikely to hold given the wide scope for model misspecification – it would be more useful to estimate the amount of bias and quantify the costs and benefits of using VA estimates for policy purposes, as in Rothstein (2015b) and Angrist et al. (2015).

References

Angrist, Joshua, Peter Hull, Parag A. Pathak, and Christopher Walters. 2015. "Leveraging Lotteries for School Value-added: Testing and Estimation." NBER Working Paper No. 21748.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." NBER Working Paper No. 20657.

Bifulco, Robert, Casey D. Cobb, and Courtney Bell. 2009. "Can Interdistrict Choice Boost Student Achievement? The Case of Connecticut's Interdistrict Magnet School Program." Educational Evaluation and Policy Analysis, 31(4): 323–345.

Cantrell, Steven, and Thomas J. Kane. 2013. "Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-year Study." MET Project Research Paper.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." American Economic Review, 104(9): 2593–2632.

Deming, David J. 2014. "Using School Choice Lotteries to Test Measures of School Effectiveness." American Economic Review, 104(5): 406–11.

Glazerman, Steven, and Ali Protik. 2015. "Validating Value-added Measures of Teacher Performance."

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Quarterly Journal of Economics, 125(1).

Rothstein, Jesse. 2015a. "Revisiting the Impacts of Teachers."

Rothstein, Jesse. 2015b. "Teacher Quality Policy When Supply Matters." American Economic Review, 105(1): 100–130.

Appendix

In this appendix, we provide further detail on how we estimate teacher value-added. The Stata code used to generate the simulated data, estimate VA, and produce the results is contained in two files: (1) simulations_table1_final.do, which generates the results in Table 1 and (2) simulations_fig1_final.do, which generates the results in Figure 1.

We estimate the value-added model in four steps:

1. Calculate student test-score gains in year $t = 2$ as $\Delta A_{i,t=2} = A_{i,t=2} - A_{i,t=1}$ and classroom-specific average gains $\Delta \bar{A}_{c,t=2} = \sum_{i \in c} \Delta A_{i,t=2}$. To simplify notation, we drop the $t = 2$ subscript for the remainder of this description.

2. Decompose the variance of test score gains into its four constituent pieces according to

$$Var(\Delta A_{i,t=2}) = Var(\psi_{st} - \psi_{s,t-1}) + Var\left(\mu_j\right) + Var(\theta_t) + Var(\varepsilon_{it} - \widetilde{\varepsilon}_{i,t-1})$$

   where $\widetilde{\varepsilon}_{i,t-1} = \varepsilon_{i,t-1} + \theta_{c(i,t-1),t-1} + \mu_{j(i,t-1)}$. We estimate these variance components as follows:

   (a) Track-Level Variance: We estimate the track-level variance component of score gains as the covariance between classroom average scores in classrooms in the same track taught by different teachers:

$$\widehat{Var(\psi_{st} - \psi_{s,t-1})} = Cov\left(\Delta \bar{A}_c, \Delta \bar{A}_{c'}\right)\Big|_{s(c)=s(c'), j(c) \neq j(c')}$$

   (b) Teacher-Level Variance: We estimate the sum of track-level and teacher-level variance as the covariance between average scores in classrooms taught by the same teacher. We then subtract the estimate of track-level variance from (a) to estimate teacher-level variance:

$$\widehat{\sigma^2_\mu} = \widehat{Var\left(\mu_j\right)} = Cov\left(\Delta \bar{A}_c, \Delta \bar{A}_{c'}\right)\Big|_{j(c)=j(c')} - \widehat{Var(\psi_{st} - \psi_{s,t-1})}$$

   (c) Individual-Level Variance: We estimate the individual level variance as the variance of test scores within classrooms, adjusted for the degrees of freedom. Note that the shock $\varepsilon_{i,t-1}$ has greater variance than ε_{it} since it includes both the lagged teacher shock and lagged classroom shock (neither of which aggregate to the classroom level because students are reshuffled across classrooms in practice):

$$\widehat{Var(\varepsilon_{it} - \widetilde{\varepsilon}_{i,t-1})} = Var\left(\Delta A_i - \Delta \bar{A}_c\right) * \frac{I}{I-1}$$

   (d) Class-Level Variance: We estimate the class-level variance as the residual variance present in the aggregate variance of test score gains after subtracting out the other three components:

$$\widehat{\sigma^2_\theta} = \widehat{Var(\theta_t)} = Var(\Delta A_i) - \widehat{Var(\psi_{st} - \psi_{s,t-1})} - \widehat{\sigma^2_\mu} - \widehat{Var(\varepsilon_{it} - \widetilde{\varepsilon}_{i,t-1})}$$

3. Calculate the shrinkage factor using the four component variances:

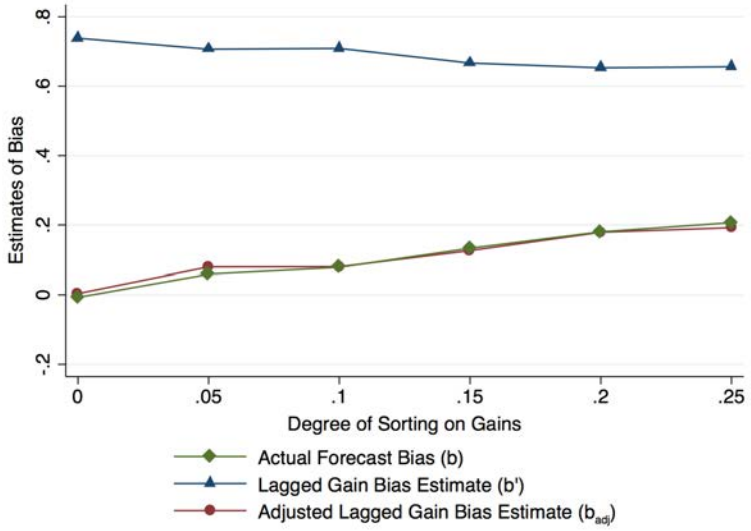
$$\lambda = \frac{\widehat{\sigma_{\mu}^2}}{\widehat{\sigma_{\mu}^2} + Var\left(\widehat{\psi_{st} - \psi}_{s,t-1}\right) + \frac{\widehat{\sigma_{\theta}^2}}{C} + \frac{Var\left(\widehat{\varepsilon_{it} - \varepsilon}_{i,t-1}\right)}{C*I}}$$

4. Estimate value-added for each teacher:

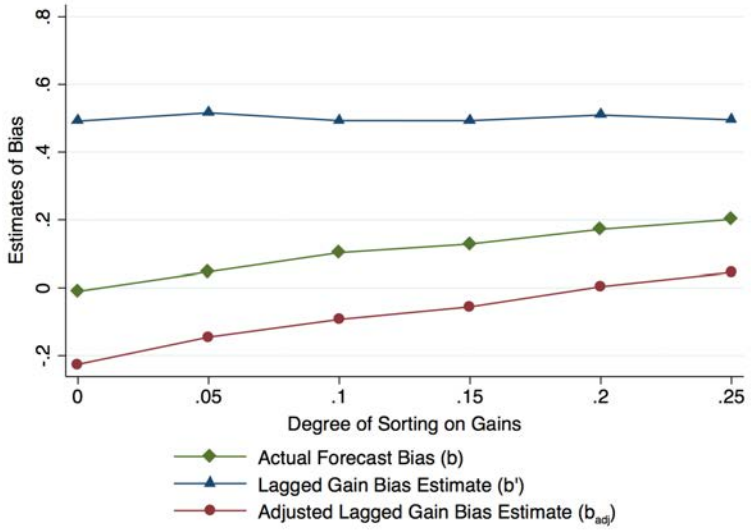
$$\hat{\mu}_j = \lambda \Delta \bar{A}_j$$

Figure 1. Estimates of Forecast Bias with Dynamic Sorting

A. Common Track Level Shocks Across Grades



B. Imperfectly Correlated Track Level Shocks Across Grades



Notes: In this figure, each point represents a coefficient from a different regression estimated using data simulated from the model described in Section 2. In Panel A, the parameter values are set as described in Appendix Table 1, except for the parameter "Degree of Sorting (Trend): Corr(α, μ)," which takes the values shown on the x-axis. In Panel B, the parameter values are as described in Panel A, except that the track level shocks are only correlated at $\rho = 0.67$ across grades within the same track. In both panels, the series "Actual Forecast Bias," is calculated as one minus the coefficient obtained from regressions of current gains on teacher VA in a randomly assigned sample of students using the specification in Column 1 of Table 1. The series "Lagged Gain Bias Estimate" shows estimates from regressions of lagged gains on VA estimates using the specification in Column 2 of Table 1. The series "Adjusted Lagged Score Bias" subtracts a correction factor to adjust for the variance of track level shocks from the "Lagged Score Bias" series, as shown in equation (6).

Table 1. Effects of Teacher Value-Added on Current and Lagged Test Score Gains

	Randomized Experiment	Lagged Scores	Lagged Scores vs. True VA	Observational Out-of-Sample Forecast	Observational, Controlling for Lagged Gain
	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	Current Gain	Lagged Gain	Lagged Gain	Current Gain	Current Gain
VA Estimate	1.010	0.709		0.991	0.833
	(0.007)	(0.013)		(0.017)	(0.016)
True VA			0.002		
			(0.004)		
Control for Lagged Gain					X
Naïve F-test for Teacher Effects		F = 2.238 p<0.001			

Notes: This table reports estimates from OLS regressions using data simulated from the model described in Section 2, with the parameter values set as described in Appendix Table 1. In columns 1-2 and 4-5, the independent variable of interest is teacher value-added, estimated using data from years t = 1 and 2 as described in the text. In Column 1, we randomly assign new students to teachers in year t = 3 and regress their test score gains on their teachers' VA estimates. Column 2 regresses lagged test score gains on teacher VA estimates in a sample of students who are assigned to teachers in year t = 3 using the same (non-random) assignment rule as in previous years. Column 3 replicates the regression in Column 2, replacing estimated teacher value-added with true teacher value-added as the independent variable. Columns 4 and 5 present "out-of-sample" estimates of forecast bias in a sample of students who are assigned to teachers in year t = 4 using the same (non-random) assignment rule as in previous years. In Column 4, we regress current test score gains on teacher VA estimates; Column 5 replicates Column 4, adding lagged gains as a control. All standard errors on regression coefficients are clustered at the track level. The F-test reported in Column 2 is implemented by regressing students' test score gains on teacher fixed effects, without clustering standard errors.

Appendix Table 1. Baseline Parameters for Monte Carlo Simulations

Parameter	Value
Number of Schools	2000
Number of Tracks per School	5
Number of Teachers per Track	4
Number of Classrooms per Teacher (C)	4
Number of Students per Classroom (I)	25
SD of Student Ability Levels (σ_δ)	0.88
SD Of Student Ability Trends (σ_α)	0.15
SD Of Teacher Value-Added (σ_μ)	0.10
SD of Classroom Shocks (σ_θ)	0.08
SD of Track Level Shocks (σ_ψ)	0.06
Correlation of Track Shocks Across Grades (ρ)	1.00
Degree of Sorting on Levels: Corr(δ,μ)	0.25
Degree of Sorting on Trends: Corr(α,μ)	0.00