Assessing the "Rothstein Test"

Does It Really Show Teacher Value-Added Models Are Biased?

Dan Goldhaber

The University of Washington

Duncan Chaplin

Mathematica Policy Research

March 2012

ABSTRACT

In a provocative and influential paper, Jesse Rothstein (2010) finds that standard value-added models (VAMs) suggest implausible and large future teacher effects on past student achievement, a finding that obviously cannot be viewed as causal. This is the basis of a falsification test (the Rothstein falsification test) that *appears* to indicate bias in VAM estimates of current teacher contributions to student learning.

Rothstein's finding is significant because there is considerable interest in using VAM teacher effect estimates for high-stakes teacher personnel policies, and the results of the Rothstein test cast considerable doubt on the notion that VAMs can be used fairly for this purpose. In this paper we show that the Rothstein test does show that students are tracked to teachers, but the tracking could be based on lagged achievement, the key control variable used in most VAMs. Our results indicate that the Rothstein test does not appear to provide additional useful guidance regarding the efficacy of VAMs, suggesting a much more encouraging picture for those wishing to use VAM teacher effect estimates for policy purposes.

I. INTRODUCTION

In a provocative and influential paper, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Jesse Rothstein (2010) reports that VAMs used to estimate the contribution individual teachers make toward student achievement fail falsification tests, which appears to suggest that VAM estimates are biased. More precisely, Rothstein shows that teachers assigned to students in the *future* have large and statistically significant estimated effects on *past* student achievement, a finding that obviously cannot be viewed as causal. Rather, the finding appears to signal that student-teacher sorting patterns in schools are not fully accounted for by the types of variables typically included in VAMs, implying a correlation between time-varying omitted variables affecting student achievement and teacher assignments. Rothstein presents this finding (his falsification test) as evidence that students are not effectively randomly assigned to teachers conditional on the covariates in the model.

Rothstein's falsification test has become a key method for academic papers to test the validity of VAM specifications. For instance, in their analysis of the impacts of teacher training, Harris and Sass (2010) report selecting school districts in which the Rothstein test does not falsify when estimating the impacts of teacher training programs. Briggs and Domingue (2011) say that they use the Rothstein test to critique value-added results that were produced for the Los Angeles public schools and later publicly released.

Koedel and Betts (2009) explore the Rothstein test more extensively. They argue that the sorting of students to teachers may be transitory in the sense that the unobserved omitted factors (influencing both achievement and the match of students and teachers) leading to positive "match shocks" in one period could be offset by negative match shocks in other periods. If so then teacher effectiveness estimates based on multiple periods of performance may be unbiased. Their empirical results appear to confirm this as they find that there is little evidence of future teacher effects, i.e. the Rothstein test fails to falsify VAMs, when teacher effect estimates are based on teachers observed in multiple classrooms (over time).

Rothstein's finding has considerable relevance because there is great interest in using VAM teacher effect estimates for policy purposes such as pay for performance (Podgursky and Springer 2007; Eckert and Dabrowski 2010) or determining which teachers maintain their eligibility to teach after some specified period of time, such as when tenure is granted (Goldhaber and Hansen 2010a; Gordon et al. 2006; Hanushek 2009). Many would argue that, if VAMs are shown to produce biased teacher effect estimates, it casts doubt upon the notion that they can be used for such high-stakes policy purposes.[1] Indeed, this is how the Rothstein findings have been interpreted. For instance, in an article published in *Education Week*, Debra Viadero (2009) interprets the Rothstein paper to

[1] Other researchers have come to somewhat different conclusions about whether VAMs are likely to produce biased estimates of teacher effectiveness. Kane and Staiger (2008), for instance, find that certain VAM specifications produce teacher effect estimates that are similar to those produced under experimental conditions, a result supported by nonexperimental findings of Chetty et al. (2011), who exploit differences in the estimated effectiveness of teachers who transition between grades or schools to test for bias. And while Koedel and Betts (2009) confirm Rothstein's basic findings about single-year teacher effect estimates, they report finding no evidence of bias based on the Rothstein falsification test when the VAM teacher effect estimates are based on teachers observed in multiple classrooms over time.

suggest that "'value-added' methods for determining the effectiveness of classroom teachers are built on some shaky assumptions and may be misleading."[2] As Rothstein (2010) himself said, the "results indicate that policies based on these VAMs will reward or punish teachers who do not deserve it and fail to reward or punish teachers who do." In a recent congressional briefing, Rothstein cited his falsification test results and said that value-added is "not fair to…special needs teachers…[or] other specialists."[3]

In this paper, we identify conditions where the Rothstein test would provide evidence that key control variables were left out of VAM models. At the same time, we show that these conditions are not plausible and that under plausible conditions, the Rothstein test will reject even when no key controls were left out of the model. We also show that under plausible conditions estimated future teacher effects are likely to be similar in size to those found by Rothstein even when the bias is very small. Thus, the Rothstein test does not appear to provide useful evidence regarding the existence or magnitude of bias. We verify these conditions theoretically and through a series of simulations. Our findings are important because, as noted above, the Rothstein test is shaping not only academic studies but also public perception about the efficacy of utilizing VAMs.

II. THE ROTHSTEIN FALSIFICATION TEST AND BIAS

A. The Value-Added Model Formulation

There is a growing body of literature that examines the implications of using VAMs in an attempt to identify causal impacts of schooling inputs and contributions of individual teachers toward student learning (for example, Chetty et al. 2011; Ballou et al. 2004; Goldhaber and Hansen 2010a; Kane and Staiger, 2008; McCaffrey et al. 2004, 2009; Rothstein, 2009, 2010; Rockoff, 2004). Researchers typically assume their data can be modeled by some variant of the following equation:

(1) $A_{ig}= \alpha+\lambda A_{i(g-1)}+\Sigma\beta_t\tau_{tig}+e_{ig}$

where A_{ig} is the achievement of student i in grade g,

α is the intercept and also the value added of the omitted teacher,[4]

λ is the impact of lagged achievement,

τ_{tig} is a dummy variable identifying if student i had teacher t in grade g,

[2] Others have cited his work in similar ways (Hanushek and Rivkin 2010; Rothman 2010; Baker et al. 2010; and Kraemer et al. 2011).

[3] Haertel et al. (2011).

[4] The intercept equals the value of the outcome when all other variables are set to 0. If the achievement scores (current and baseline) are mean centered, the intercept equals the outcome for a student with an average baseline score who has the omitted teacher.

β_t is the impact of teacher t compared to the omitted teacher,[5] and

e_{ig} is an error term that represents other factors that affect student learning.

If e_{ig} is uncorrelated with the other variables in equation 1, then the impact of teacher t can be estimated by regressing student achievement (A_{ig}) on prior student achievement ($A_{i(g-1)}$) and dummy variables identifying the teacher student i had in grade g ($\tau_{t,i,g}$).[6] Of course, this model makes a number of assumptions about the nature of student learning; see, for instance, Harris et al. (2010), Rothstein (2010), or Todd and Wolpin (2003) for more background on these assumptions.[7]

Rothstein implements his falsification test for several VAM specifications; equation (1) is similar to the "VAM2" model discussed by Rothstein (2010, p. 182) in that it also allows the coefficient on lagged achievement to differ from 1 and excludes student-fixed effects.[8] We focus on this model specification because it is a commonly used approach to assess teacher effectiveness, and, more importantly, has been found to yield estimates of effectiveness that are consistent with those estimated when teachers are randomly assigned to classrooms (Kane and Staiger, 2008).[9]

Rothstein questions whether standard VAMs produce unbiased estimates of teacher effect on student learning.[10] In particular, he describes a number of ways in which the processes governing the assignment of students to teachers may lead to erroneous conclusions about teacher effectiveness. Of particular concern is the possibility that students may be tracked into particular classrooms based on skills that are not accounted for by $A_{i(g-1)}$.[11]

[5] In the rest of this paper, we generally refer to β_t as the impact of teacher t. This can be thought of as equivalent to saying that the results are normalized so that the impact of the omitted teacher is 0.

[6] Like much of the value-added literature, Rothstein does not try to separate classroom and teacher effects.

[7] Note that, even if VAMs produce effect estimates that are unbiased, they may not be very reliable. For more on this, see Goldhaber and Hansen (2010b), McCaffrey et al. (2009), and Schochet and Chiang (2010).

[8] Value-added models often also include school and grade fixed effects and a vector of student and family background characteristics (for example, age, disability, English language status, free or reduced-price lunch status, race, ethnicity, whether a student has previously been retained in grade, and parental education).

[9] More precisely, Kane and Staiger fail to reject the null hypothesis of the equivalence of differences between pairs of teachers under nonexperimental conditions (i.e. when tracking is possible), and another period when they are randomly assigned to classrooms. This was not true of VAMs utilizing student gains as the dependent variable or models that included student fixed effects.

[10] Parameter estimates from many statistical models can often be consistently estimated with large sample sizes but remain biased when sample sizes are small relative to the number of parameters being estimated. Kinsler (2011) investigates how small sample size issues affect the results of the Rothstein test.

[11] This could happen for a number of reasons. For example, principals might assign students to teachers based in part on unobserved factors that impact current achievement but are not captured by lagged achievement. Principals might do this either because they believe that certain teachers have a comparative advantage in educating certain students or because the principals are rewarding certain teachers with choice classroom assignments. Similar results could hold if parents lobby to have their children assigned to particular teachers.

B. Bias in VAM Teacher Effect Estimates

If equation 1 is estimated using ordinary least squares, the estimated impacts of a teacher can be biased for a number of reasons. To derive a formula for this bias, we divide the error term (e_{ig}) into two components—ov_{ig}, which, *if it exists*, we assume to be correlated with at least some of the covariates included in the model even after controlling for the others, and eu_{ig}, which is uncorrelated with any of the covariates in the model.

Thus,

$$e_{ig} = \gamma' ov_{ig} + eu_{ig}$$

where γ is the coefficient on ov_{ig}. γ is also the coefficient that would be obtained on ov_{ig}, were it added to equation 1.

It should be noted that, by definition, if ov_{ig} exists it would cause bias for at least some coefficient estimates. However, as we describe below, it can exist and not necessarily cause bias for the estimated teacher effects.

The general formula for omitted variable bias for the effect of teacher t takes the form:

$$Bias(\hat{\beta}_t) = E(\hat{\beta}_t) - \beta_t = \gamma\pi_{te}$$

where $\hat{\beta}_t$ = estimate of β_t and π_{te} = coefficient on τ_{tig} from a regression of ov_{ig} on all right-hand side variables in equation 1 (except e_{ig}).

It can be shown that,

$$(2)^{12} \qquad \pi_{te} = cov(ov^*_{ig}, \tau^*_{tig})/V(\tau^*_{tig})$$

where ov^*_{ig} and τ^*_{tig} are the residual values of ov_{ig} and τ_{tig} that remain after controlling for other variables in a linear regression model.[13]

This formulation is helpful because it shows that only the residual value of ov_{ig} matters for bias in the estimated teacher effects. In other words, ov_{ig} would not cause bias in the estimated teacher effects if its residual is uncorrelated with the residual of τ_{tig}.[14]

[12] Equation 2 holds because each coefficient estimate from any linear regression can be obtained by regressing the residualized outcome on the residualized version of the corresponding right-hand side variable in that equation. Each residualized variable is equal to the residual obtained after regressing the original variable on all of the other right-hand side variables in the equation (Goldberger 1991).

[13] We use the abbreviation "cov" for "covariance" and "var" for "variance" in equations throughout the paper. We use "*" to indicate a residual from a linear regression of the variable in question on lagged achievement. We could also write $ov_{ig} = \pi_{xe}\, x_{itg} + ov_{ig*,}$ where x_{itg} represents all the covariates in the model except τ_{tig} and π_{xe} is a vector of coefficients on those variables. In our example, the vector x_{itg} includes lagged achievement and all other teacher variables in equation 1.

If one estimates a linear VAM, then the residual of ov_{ig} would also be based on a linear regression. Consequently, another way to state the condition of obtaining unbiased teacher effects in the presence of an omitted variable in a linear VAM is as follows: if the omitted variable is a linear function of lagged achievement and the other control variables in the equation, then it need not cause bias in the estimated teacher effects because the partial correlation is picked up by the included variables.

If the omitted variable is a nonlinear function of lagged achievement, then it is more likely to cause bias. However, even in this case, it may still be possible to obtain unbiased estimates of the teacher effects if the omitted variable is time-invariant (that is, $ov_{ig}=ov_i$). In this situation, it may be possible to address the issue of tracking based on time-invariant factors through the inclusion of a student fixed effect, which captures time-invariant student or family background attributes that affect current achievement in ways not captured by previous achievement.[15] But one of the significant contributions of Rothstein's work is that he raises an additional concern about VAMs: he postulates that student sorting into classrooms is "dynamic" in the sense that an omitted factor that affects the error term may be time-varying and correlated with placement into classrooms (that is, a form of tracking). This could lead to bias. He specifically suggests that the omitted factors cause the errors to be negatively correlated over time.[16] This could be due to compensating behavior whereby students who have a good year (along unobserved lines) receive fewer inputs in the following year than students who are otherwise similar.[17] This dynamic form of tracking cannot be accounted for by the inclusion of a simple student fixed effect.

C. Rothstein's Falsification Test

Rothstein's falsification test relies on the notion that evidence of the statistical significance of *future* teachers in predicting *past* achievement suggests a misspecification of the VAM (Rothstein 2010). Similar falsification tests are often used in the economics literature (Heckman and Hotz 1989; Ashenfelter 1978). However, these tests are generally used to check the specification of models that differ in important ways from VAMs. For example, Heckman and Hotz (1989), in the context of studying the impacts of a job training program, propose a general falsification test for nonexperimental estimators based on the same underlying concept as Rothstein—that a treatment cannot affect past outcomes. And Rothstein (2010) cites Ashenfelter (1978), which also focused studying job training. But the analogy with falsification tests used in the job training literature may

(continued)

[14] The variable ov_{ig} could still cause bias in the estimated impacts of other covariates, like the lagged achievement variable, and therefore be considered an omitted variable for the regression.

[15] Studies differ on the efficacy of including student fixed effects—Harris and Sass (2010) argue for it, while Kane and Staiger (2008) argue against. Koedel and Betts (2009) find that student fixed effects are statistically insignificant in the models they estimate.

[16] Rothstein's evidence of a negative correlation of errors over time is derived from a VAM without student fixed effects. If important time-invariant omitted student factors exist, implying the need for student fixed effects, we would expect to see a positive correlation across grade levels between the errors in a model estimated without student fixed effects. Given this, it seems unlikely that student fixed effects explain Rothstein's finding of bias.

[17] This could happen if family contributions to achievement are negatively correlated over time, conditional on past achievement. Alternatively, this could happen if the impacts of the error term decay at a different rate than other factors. By allowing the coefficient on lagged achievement to be less than one, the VAM we consider allows for decay, but implicitly assumes that the decay is the same for all factors (prior achievement, teachers, and the error term).

not hold because there is no reason to believe that there is a nonlinear relationship between lagged earnings and double-lagged earnings as is likely to be the case with lagged and double lagged achievement.[18]

We focus our paper on the version of the Rothstein test described in footnote 13 of his paper, where Rothstein states that "…random assignment conditional on A_{ig-1} will be rejected if the grade-g classroom teachers predict A_{ig-2} conditional on A_{ig-1}" (this is described more formally in the next subsection).

It is important to note that the outcome in this equation is grade *g-2* achievement. If one were to instead look at the impact of grade g teachers on grade g-1 achievement and control for grade g-2 achievement, then the nature of the test changes, although the same general properties hold. In particular, the test will often reject when there is no bias for estimated teacher effects, as discussed below.

Rothstein implements his falsification test using a linear regression. Specifically, achievement in grade 3 is regressed on grade 4 achievement and grade 5 teachers. The test is whether the coefficients on the grade 5 teachers are jointly significant.

Rejecting the null of no future teacher effects might then be taken as evidence of bias. In other words, many people may assume it to imply a correlation between the independent variation in teacher assignments and the error in the achievement equation. But a more precise formulation of the necessary exclusion restriction for unbiased teacher effects is that current grade teacher assignments are orthogonal to all residual determinants of the *current* score that are not accounted for by the exogenous variables included as controls in the model. This does not, however, mean that *future* grade teacher assignments need necessarily be orthogonal to the residual determinants of past scores. Rothstein's falsification test does provide evidence of tracking and that tracking *could* cause VAM misspecification. But, as we show below, the Rothstein test will reject, even in the absence of any omitted variables that might lead to biased teacher effect estimates.

D. Comparing the Rothstein Falsification Test to a Formal Condition for Bias

The central connection between the Rothstein falsification test and a formal condition for bias is through the tracking of students. However, tracking of students to future teachers (implied by the Rothstein test) need not imply that students are tracked to current teachers in the same way.[19] For example, students can be tracked into future teacher classrooms based, at least in part, on past achievement, even if they were randomly assigned to current teachers. For this reason, the Rothstein test only applies to bias for current teacher effect estimates if tracking systems are consistent across

[18] This implies that while the results we present below suggest concern regarding the use of falsification tests for a VAM, they do not rule out the use of falsification tests in other situations, such as those considered by Heckman and Hotz (1989), and even in other education research in which the goal is to estimate effects of programs or policies that target large numbers of classrooms.

[19] Koedel and Betts (2009) argue that using multiple cohorts of students to inform on teacher effect estimates helps to mitigate bias because unobserved student-teacher sorting is transitory.

grades.[20] Consistent tracking across grades also means that we can lag the Rothstein test one period and get the same results. We do this because it provides a means of concisely comparing the Rothstein falsification and bias tests.

For simplicity, in comparing the Rothstein and bias tests in this section of our paper we utilize a simple VAM with only one school and two teachers in each grade (in Section III we show simulations with more complex data structures). The dummy for one teacher is omitted from the regression and the current grade teacher assignment depends entirely on lagged achievement. Thus:

(3)[21] $A_{ig}= \lambda A_{i(g-1)} + \beta_{1g}\tau_{1ig}+e_{ig}$

where $\text{cov}(e_{ig},A_{i(g-1)}) = \text{cov}(e_{ig},\tau_{1ig}) = 0$.

We specify a flexible functional form for tracking for teacher 1 (the one with higher value added):

$\tau_{1ig} = T(A_{i(g-1)})$[22]

Rothstein's falsification test is based on a regression of lagged achievement on current achievement and future teachers. In Table IV of his paper, he uses a similar test in which he adds in current teachers as an additional set of control variables in the model estimating future teacher effects. Our conclusions about the relationship between the Rothstein falsification and formal condition for bias are unaffected by the version of the Rothstein test employed. More precisely, as discussed in Section III, simulations that use this second test will also often reject when there is no bias and when students are randomly assigned conditional on the covariates in the model. Thus, the issues we raise below about the relationship between the Rothstein test and the conditions for bias are pervasive regardless of how the test is implemented.

As noted earlier, to simplify our discussion we lag the Rothstein test one period so that instead of focusing on the relationship between future teachers and lagged achievement, the test is based on the relationship between current teachers and double-lagged achievement. This enables us to compare the conditions for bias and the Rothstein test using the same set of teachers. Again, if the possible sources of bias are the same across grades, this simplification has no effect on our results.

The Rothstein test using future teachers can be written as:

$A_{i(g-1)} = \lambda 1 A_{ig} + R_{1\ (g+1)}\tau_{1i(g+1)}+w_{i(g-1)}$

[20] Likewise, the implicit assumption is that tracking policies are consistent over time. If tracking does vary across grades then the Rothstein test might best be used to try to identify the potential for bias in estimated effects of future teachers on future student achievement (i.e. grade 5 teachers on grade 5 achievement scores in our case).

[21] To simplify our presentation, we omit intercepts and measurement error from these equations. Adding them back does not change our substantive findings. Numerous methods exist to obtain unbiased estimates in the presence of such measurement error (Potamites et al. 2009; Meyer 1999; Fuller 1987).

[22] $T()$ is bounded between 0 and 1 and $dT()/dA_{(i,g-1)} > 0$.

where $R_{1(g+1)}$ describes the regression-adjusted relationship between future teachers and lagged achievement.

When we lag this one period, we get,

(4)[23] $A_{i(g-2)} = \lambda 1 A_{i(g-1)} + R_{1g}\tau_{1ig} + w_{i(g-2)}$

The Rothstein test involves estimating whether or not R_{1g} differs from 0.

The numerator in R_{1g} is the covariance between current teachers and the residual from a regression of double-lagged achievement on current achievement.[24]

(5) $A_{i(g-2)} = \lambda 2\ A_{i(g-1)} + u_{i(g-2)}$

If R_{1g} is 0 then $\lambda 2$ equals $\lambda 1$ and $u_{i(g-2)}$ equals $w_{i(g-2)}$. Thus, one can test to see if R_{1g} differs from 0 using the following covariance:

$\text{cov}(\tau_{1ig}, u_{i(g-2)} | A_{i(g-1)}) <> 0$[25]

For comparison, here is the condition for obtaining biased estimates:

$\text{cov}(\tau_{1ig}, e_{ig} | A_{i(g-1)}) <> 0$

Formulated in this way, the Rothstein falsification test and formal condition for bias appear quite similar but differ in a key respect—e_{ig} is not the same as $u_{i(g-2)}$. One can, therefore, generate data in which there is no bias but the Rothstein test rejects (and vice versa).

We break up the error term from equation 5 into three pieces to show factors that might cause the Rothstein test to falsify unbiased VAMs. Rearranging and substituting out for $A_{i(g-1)}$ yields:

$$u_{i(g-2)} = A_{i(g-2)} - \lambda 2(\lambda A_{i(g-2)} + \beta 1_{(g-1)}\tau_{1i(g-1)} + e_{i(g-1)})$$

(6) $$= (1 - \lambda 2\lambda)\ A_{i(g-2)} - \lambda 2\beta 1_{(g-1)}\tau_{1i(g-1)} - \lambda 2\ e_{i(g-1)}$$ [26]

Equation 6 shows that $u_{i(g-2)}$ can be written as a linear function of three variables—$A_{i(g-2)}$, $\tau_{1i(g-1)}$, and $e_{i(g-1)}$.The Rothstein test is based on testing whether $\text{cov}(\tau_{1ig}, u_{i(g-2)} | A_{i(g-1)}) <> 0$, which means it will reject if any of the following three conditions hold:

[23] As shown earlier, footnote 13 of Rothstein's paper uses the same grade levels—g, g-1, and g-2.

[24] More precisely, $R_{1g} = \text{cov}(\tau_{1,i,g}, A_{i(g-2)} | A_{i(g-1)}) / \text{var}(\tau_{1,i,g} | A_{i(g-1)}) = \text{cov}(\tau_{1,i,g}, u_{i(g-2)} | A_{i(g-1)}) / \text{var}(\tau_{1,i,g} | A_{i(g-1)})$ since $u_{i(g-2)}$ is the residual that remains after regressing $A_{i(g-2)}$ on $A_{i(g-1)}$.

[25] Throughout, we use $\text{cov}(x,y | A_{i(g-1)})$ to describe the covariance that remains between variables x and y after controlling for $A_{i(g-1)}$ using a linear regression. The conditioning does not change this covariance, but we use the conditional notation here to help stress how similar this covariance is to the covariance on the next line.

[26] This equation is similar to Rothstein's equation 7 except that it includes double-lagged achievement on the right-hand side.

(C1) $\text{cov}(\tau_{1ig}, A_{i(g-2)} \mid A_{i(g-1)}) <> 0$

(C2) $\text{cov}(\tau_{1ig}, \tau_{1i(g-1)} \mid A_{i(g-1)}) <> 0$

(C3) $\text{cov}(\tau_{1ig}, e_{i(g-1)} \mid A_{i(g-1)}) <> 0$

E. Exploring the Conditions That Cause the Rothstein Test to Falsify

The first condition that would cause the Rothstein test to reject is that current teachers are conditionally correlated with double-lagged achievement, after controlling for lagged achievement in a linear regression. This is not implausible, as school systems may not have ready access to lagged achievement scores at the point at which teacher assignment decisions are made.[27] Consequently, many schools may use double-lagged achievement for tracking decisions.

The second condition that would cause the Rothstein test to reject occurs when current teachers are conditionally correlated with lagged teachers, after controlling for lagged achievement. This is likely if some classrooms disproportionately consist of students who shared the same classroom in the previous year, perhaps because schools intentionally keep certain students together (or apart). We refer to this as "classroom tracking."[28] While this type of tracking may be prevalent in some situations, Rothstein presents evidence suggesting that the amount of movement between classrooms across years is large enough to suggest that this type of tracking was not common in the data he analyzed.[29]

Conditions 1 and 2 both relate to the possibility that there is a variable left out of the VAM that affects tracking and that might, therefore, cause bias. However, if these variables (double-lagged achievement and lagged teachers) affect current achievement only through their impacts on lagged achievement, as is implied by equation 3, then they may cause the Rothstein test to reject, but their omission from the VAM will cause no bias.

The third condition that would cause the Rothstein test to reject occurs when the current teacher is conditionally correlated with the lagged error term, after controlling for lagged achievement. This condition might seem unlikely to matter since $e_{i(g-1)}$ is enters into the equation for $A_{i(g-1)}$ directly. One might therefore assume that controlling for $A_{i(g-1)}$ would account for a correlation between $e_{i(g-1)}$ and $\tau_{1,i,g}$. This, however, turns out not to be the case because the falsification test is linear while both the current teacher and double-lagged achievement are nonlinear functions of lagged achievement and the nonlinearities can be correlated. In fact *teacher assignments are almost certain*

[27] For instance, Mathematica does value-added work in various states and localities where teacher effect estimates are needed in a timely way to inform key policy decisions. In many of these locations, state achievement score data from the spring of one school year are often not available until the fall of the following school year, too late to affect tracking decisions for that year. See, for example, Potamites et al. (2009) and Chaplin et al. (2009).

[28] There are alternative tests for bias caused by these types of tracking. For example, for the first condition, one can include double-lagged achievement in the VAM model and test to see if the estimated coefficient estimates on current teachers change compared to a model without that variable (Rothstein 2009). Similarly, for the second condition, one can add lagged teachers to a standard VAM. Results of this later test are likely to be very imprecise for many teachers, especially in smaller schools, if most of their students come from a single lagged teacher.

[29] Rothstein, 2010, page 193.

to be a nonlinear function of lagged achievement if lagged achievement (or a latent variable that is correlated with it) is used for tracking.[30] Consequently, the linearly based falsification test can suggest implausible current teacher impacts on double-lagged achievement in an unbiased VAM. We describe this issue in detail in Appendix A and show evidence of this in our simulations.

This third condition is important because Rothstein (2010) expresses particular interest in the distribution of error terms. In particular, he acknowledges that evidence that future teachers are statistically significant predictors of past achievement is not itself proof of bias for current teachers.[31] He goes on to say that tracking accompanied by negative correlation in the errors across grades (that is, $\text{cov}(e_{ig}, e_{(i,g-1)}) < 0$) "strongly suggests" bias for current grade teachers. More precisely he says,

> "A correlation between treatment and some pre-assignment variable X need not indicate bias in the estimated treatment effect if X is uncorrelated with the outcome variable of interest. But outcomes are typically correlated within individuals over time, so an association between treatment and the lagged outcome strongly suggests that the treatment is not exogenous with respect to post-treatment outcomes (Rothstein 2010)."

We agree that, if the errors are correlated (either negatively or positively) and there is tracking, then the teacher effects will probably be biased (see Appendix B for more detail on this point) and the Rothstein falsification test will reject based on condition 3 (see Appendix A). However, as we show, the test will also reject based on conditions 1, 2, or 3 *even without negatively correlated errors*. Thus, one cannot use the test to definitively identify bias caused by negatively correlated errors. Similarly, one cannot use the Rothstein test to definitively identify variables left out of the VAM that might cause bias (such as double-lagged achievement or lagged teachers) since the test will reject based on condition 3 even without conditions 1 or 2.

As discussed above, Condition 3 is a key part of our paper. Hence, we provide a brief explanation of why it matters. First, note that one can think of a VAM model as being implicitly based on a series of simultaneous equations—one for the current test scores, and one tracking equation for each teacher. The tracking equations are inherently non-linear since the teacher outcome variables are necessarily bounded between 0 and 1 (and, in many VAMs are modeled as simple binary variables). In this context, the Rothstein tests can be seen as "reverse" regression versions of the tracking equations. More precisely, in the tracking equations current teachers depend on lagged test scores and possibly other variables, like double-lagged test scores and lagged teachers. In the Rothstein tests right-hand side variables from the tracking equation are used as the outcomes and the outcomes from the tracking equations (the teacher variables) are used as right-hand side variables. Since the original tracking equations are non-linear, the "reverse regressions" are as well.

[30] Current teachers are a nonlinear function of lagged achievement because the tracking equation is bounded between 0 and 1. Double-lagged achievement can be described using a nonlinear function of lagged achievement because the lagged teachers create discontinuous jumps in lagged achievement that are not in double-lagged achievement. The two sources of nonlinearity can be correlated because both depend on lagged achievement. This result suggests that the Rothstein falsification test may not work well for VAM..

[31] More generally, in personal correspondence with others (see Chetty et al. 2011, footnote 53), Rothstein has stated that his test is "neither necessary nor sufficient for there to be bias in a VA estimate." Rather, his test suggests cause for concern about bias that might be caused by unobservables. We view our findings as showing conditions under which that bias might also be small. Chetty et al. (2011) present nonexperimental evidence suggesting small bias caused by unobservables, supporting earlier experimental findings by Kane and Staiger (2008).

However, since the Rothstein tests use linear specifications, they can reject because of these non-linearities as we illustrate below.

III. SIMULATION RESULTS

We performed a number of simulations that illustrate the findings reported in the preceding section. For consistency with Rothstein (2010), we simulated data for grades 3 through 5.[32] We conducted several falsification tests along with a formal test for bias for the estimated grade 5 teacher effects.[33]

As noted above, Rothstein's falsification test is based on a regression of lagged achievement on current achievement and future teachers, but he also implements a variant of the falsification test in which he adds in current teachers as an additional set of control variables in the model estimating future teacher effects. There also may be room for confusion about the way to implement the falsification test since Rothstein's findings in popular press have been described as asking whether models "show that 5th grade teachers have effects on their students' test scores in 3rd and 4th grades" (e.g. Viadero, 2009). Given this, we implement three different versions of the falsification test: we test whether the estimated "future teacher effects" differs from 0 in a specification consistent with equation 4 above, whether they differ from 0 in a variant of this that includes current teachers, and, finally whether they differs from 0 in a model similar to the first except that the grade 3 and 4 achievement scores are switched so that grade 4 becomes the outcome and grade 3 is on the right hand side. Following are equations for each of the tests we conduct.

Rothstein 1) $A_{i(g-2)} = \alpha_{r1} + \lambda_{r1} A_{i(g-1)} + \Sigma R1_t \tau_{tig} + \Sigma S1_s s_{isg} + w1_{i(g-2)}$

Rothstein 2) $A_{i(g-2)} = \alpha_{r2} + \lambda_{r2} A_{i(g-1)} + \Sigma R2_t \tau_{tig} + \Sigma S2_s s_{isg} + \Sigma \beta 2_t \tau_{ti(g-1)} + w2_{i(g-2)}$

Rothstein 3) $A_{i(g-1)} = \alpha_{r3} + \lambda_{r3} A_{i(g-2)} + \Sigma R3_t \tau_{tig} + \Sigma S3_s s_{isg} + w3_{i(g-1)}$

where the s_{isg} variables are dummies identifying which school student i attended in grade g, the Sm parameters are the "effects" of these schools in test m, the λ_{rm} are the coefficients on achievement, the α_{rm} are the intercepts, and the $wm_{i(g-2)}$ are the errors. One school is omitted so we can include an intercept and one teacher per school is omitted so we can include the school effects.

In each case the parameters of interest are the "future" teacher effects described by the vectors R1, R2, and R3.

[32] We started with normally distributed achievement in grade 2 and then added in teacher effects and normally distributed errors for achievement in grades 3, 4, and 5.

[33] The grade 5 teachers can be thought of as future teachers in grade 4 using the regular Rothstein test or current teachers using our revised Rothstein test in which all variables are lagged one period. As noted earlier, if tracking systems are stable across grades, then the choice of grade levels will not matter.

For parsimony's sake the results we describe below are for the first of these specifications (the one that is consistent with footnote 13 in Rothstein's paper, but we report the findings from simulations using the two alternative specifications in Appendix C. And, importantly, we find in the simulations that *the primary results hold up regardless of the version of the falsification test employed: the test will reject VAMs in cases where there is no bias and fail to reject in cases where there is bias.* Thus, the key findings are not sensitive in substantively important ways to the precise formulation of the falsification test.

We assume the errors in the achievement equations are jointly normally distributed and have standard deviations of 0.4. We set the coefficient on lagged achievement to either 0.91 or 0.95, depending on the model, to keep the achievement level standard deviations close to 1.[34] The standard deviations of grade 5 teachers were set to 0.1. The standard deviations for grades 3 and 4 teachers were set to either 0.1 or 0, depending on the model.

We simulated data for 200 schools with four teachers per school and 20 students per teacher for a total of 800 teachers and 16,000 students in each model.

We simulated data setting the school effects to zero but controlled for school effects both in our VAMs and in the Rothstein tests. Thus, we estimated effects for three teachers in each school, or a total of 600 teachers across the sample.

We based tracking on various subsets of the following five factors: (1) previous achievement, (2) double-lagged achievement, (3) the previous teacher (a dummy variable), (4) a random component that has a standard deviation of 0.2,[35] and (5) an omitted variable (discussed below) with a standard deviation of 0.2. Within schools, we split students into four groups of 20 each based on an indicator variable equal to the sum of the tracking factors used. The factors used vary depending on the model, as described in Table 1. In some models, the achievement error terms have a negative correlation of around -0.25.[36]

Rothstein (2010) finds almost no correlation in errors across two periods. Rather, he only finds correlations in errors across contiguous periods. To be consistent with his evidence, we generated errors with these properties (correlations between contiguous periods but no correlation across non-contiguous periods). More precisely, we generated errors using the following formula:

$$e_{ig} = \omega_1 u_{ig} + \omega_2 u_{i(g-1)}$$

[34] The choice of the coefficient on lagged achievement does not impact our substantive findings as long as it is not zero. The achievement scores all have standard deviations between 0.98 and 1.01.

[35] The standard deviation of 0.2 for the tracking error implies a high degree of tracking since lagged achievement has a standard deviation of around 1.0. We use a high degree of tracking because this increases the chances of finding evidence of bias. Indeed, in the extreme, if the random tracking error had a very large standard deviation we would approach random assignment of students to teachers and get almost no bias. Nevertheless, as a robustness check we also simulate data (discussed below) with much larger standard deviations for the random error term, and thus, less tracking.

[36] Rothstein simulates data in Appendix C of his paper using a negative correlation of -0.25. He reports correlations of −.21 for math and −.19 for reading for the residuals from the VAM based on the North Carolina data he analyzes. These estimates are not adjusted for measurement error. He does report that the correlations are too large to be caused by measurement error for his "VAM1" model, which assumes a coefficient of one on lagged achievement.

where w_1 and w_2 are weights chosen to generate errors with the specified variances and correlations across grades. The u_{ig} variables are uncorrelated across grades so the errors separated by two or more periods are also uncorrelated.

As noted above, some models include an omitted variable that impacts tracking decisions. That variable also impacts current achievement scores. It is not correlated with lagged achievement.

For each model, we tested to see if the model was rejected using the Rothstein falsification test and also if the impact estimates for grade 5 teachers were biased.[37] To test for bias, we did a joint test of the difference of each of the teacher effect estimates from their true values, which were used to simulate the data (allowing for correlations across the estimates). We describe the magnitude of the estimated future teacher effects from the Rothstein test using their standard deviations. We show correlations between the estimated teacher effects and the true effects as a way of assessing the magnitude of the bias.

We present five sets of results in Table 1. The first set of columns (under "Results by Condition") demonstrates how the Rothstein test performed based on the three conditions discussed above. These are estimated without negatively correlated errors to show that the test will reject even under those conditions.

The second set of columns (under "Linear Falsification Test") covers findings with grade 4 achievement linearly associated with grade 3 achievement; the Rothstein test "works" in this situation to identify variables left out of the VAM (double-lagged achievement and lagged teachers), but as we explain this is not a plausible data generation process.

The third set of columns (under "Negatively Correlated Errors") presents results based on the most plausible data generation conditions—when the errors are negatively correlated (as Rothstein found) and when the grade 3 and 4 achievement scores are not linearly related. Under these conditions, the Rothstein test worked in the sense that it falsified the model when the VAM produced biased teacher effect estimates. However, as we explain below, the test still can't be used by itself to identify this type of bias.

The fourth set of columns (under "Failing to Falsify") presents cases in which the Rothstein test failed to falsify when it should have done so, a point that Rothstein acknowledges in his paper. The last column (under "Rejecting RA") shows that the Rothstein test can reject in spite of random assignment (RA) of teachers to classrooms.

The first four rows show the parameters used to generate the data for each model. Blanks indicate zero values. The last six rows of Table 1 show our results. The bias test is a joint test of the difference between the estimated teacher effects and the true teacher effects from equation 1 above, accounting for the fact that one teacher in each school was dropped. The Rothstein test is a joint test of the significance of current teachers for predicting double-lagged achievement controlling for lagged achievement (like equation 4 above but with multiple teachers and with school effects). The standard deviation of the VAM estimates comes from the VAM results. The standard deviation of the Rothstein estimates comes from the coefficients on the teacher dummies produced in the

[37] To estimate the standard deviation of current teacher effect estimates, we use estimates based on equation 1.

Rothstein test. The raw correlation between the true and estimated teacher effects is Raw cor $(\beta_t, \hat{\beta}_t)$ and Adjusted cor $(\beta_t, \hat{\beta}_t)$ is the correlation adjusted for estimation error.

Table 1. Simulation Results for Rothstein Falsification Test and Bias Test, by Model

Parameter	Results by Condition			Linear Falsification Test			Negatively Correlated Errors			Failing to Falsify		Rejecting RA
Result	1	2	3	4	5	6	7	8	9	10	11	12
Teachers Tracked on	A_4, A_3	A_4,τ_4	A_4	A_4,A_3	A_4,τ_4	A_4	Rand	A_4	A_4	ov^*_g	$ov^*_{g,}$,A_4	τ_4
Cor(e_g,e_{g-1})= Cor(e_{g-1},e_{g-2})							-0.25	-0.25	-0.25		-0.25	
Var(ov^*_g)										0.2	0.2	
Std Dev(β_3, β_4)	0.1	0.1	0.1				0.1	0.1		0.1		0.1
Bias Test	0.95	0.95	0.98	1.01	0.96	0.92	1.01	1.00	0.88	**3.34***	**2.02***	1.02
Rothstein Test	**8.34***	**8.91***	**1.27***	**7.11***	**7.97***	1.03	0.99	**1.18***	0.96	1.00	0.97	**2.01***
Std Dev VAM Estimates	0.132	0.135	0.137	0.132	0.138	0.129	0.138	0.135	0.133	0.189	0.230	0.133
Std Dev Rothstein Estimates	0.331	0.292	0.149	0.311	0.278	0.130	0.129	0.144	0.125	0.133	0.126	0.174
Raw cor(β_t, $\hat{\beta}_t$)	0.71	0.76	0.75	0.71	0.74	0.71	0.73	0.73	0.75	0.52	0.42	0.71
Adjusted cor(β_t, $\hat{\beta}_t$)	0.99	1.00	1.00	0.98	0.98	0.99	0.97	0.99	1.00	0.60	0.47	0.97

Notes: A_g is achievement in grade g, τ_g is a vector of teacher dummies, ov^*_g is an omitted variable that causes bias, e_g is an error term, and β_g is a vector of coefficients on the teacher dummies. Blanks indicate zeros. Tracking is based on the sum of the variables indicated in the table and a random error with standard deviation of 0.2 Rand means this random error was the only one used for tracking. School effects are set to zero. Achievement has a standard deviation close to one. The errors e_g, e_{g-1} and e_{g-2} are in the achievement equations and have standard deviations of 0.4. The variable ov^*_g affects both achievement and tracking in grade 5, but is uncorrelated with previous grade variables. Lambda is 0.91 in all models except for 7, 8, 9, and 11, where it is 0.95. The bias test is a joint test of the difference between the estimated and true teacher effects in a standard VAM. The Rothstein test is described in the text. The test statistics are F-tests. The cut-points for 5 and 1 percent statistical significance given our sample sizes are 1.10 and 1.14 respectively. Test statistics in bold with a "*"are significant at the 5 percent level. The standard deviation of grade 5 teacher effects is 0.1 in all models. Std Dev(β_3, β_4) refers to the standard deviations of the teacher effects in grades 3 and 4. Each model has 200 schools, four teachers per school, and 20 students per teacher. The regressions control for school effects so we only estimate teacher effects for 600 teachers in each model (three teachers per school and 200 schools).

Adjusted cor(β_t, $\hat{\beta}_t$) is an estimate of the correlation that would be observed between β_t and $\hat{\beta}_t$ in the absence of any estimation error (Spearman 1904; Goldhaber and Hansen 2010b).

The first three columns show cases in which the Rothstein test rejected but there was no bias, relating them to the three conditions discussed above. For the first two conditions, the reason for the Rothstein test to reject seems fairly clear—current teachers were selected in part based on double-lagged achievement and/or lagged teachers, both of which are components of the error term from the Rothstein test. Hence, current teachers are correlated with that error. In the third column, however, neither condition was present and yet the Rothstein test still rejects. Here the false rejection was caused by the fact that the bivariate relationship between double-lagged achievement and lagged achievement is nonlinear, whereas the Rothstein falsification test is linear, as discussed earlier. Appendix A explains in more detail why this outcome is possible.

One of Rothstein's findings is that the magnitudes of the future teacher effects are quite large. As shown in columns 1 through 3, we found this result in models without bias. Indeed, even in column 3, when rejection is due only to the nonlinearity issue, the estimated future teacher effects from the Rothstein test are still about the same size as the estimated current teacher effects from the VAM.

The fact that the grade 5 future teacher effects are noticeable for conditions 1 and 2 might not seem surprising, given that those conditions imply that variables that affected tracking were left out of the VAM equation. Perhaps more surprising is the fact that the nonlinearities cause such a large future teacher coefficient estimate for condition 3. One reason for this is that the denominator of the coefficient estimate on the grade 5 teacher is the variance in the grade 5 teacher that remains after controlling for lagged achievement. If lagged achievement predicts the current teacher well in a linear model, then the denominator (residual variance) may be small, resulting in a relatively large grade 5 teacher coefficient estimate. This means that the "future teacher effects" produced by the Rothstein test have magnitudes similar to true teacher effects in models with little or no bias. Thus, their magnitudes could be misleading in regard to the magnitude of the bias identified.

In columns 4, 5, and 6 (under "Linear Falsification Test") we show conditions under which the Rothstein test can be used to identify variables left out of a VAM—in particular when the grade 3 and 4 achievement levels are linearly related. In columns 4 and 5 the Rothstein test still rejects suggesting that tracking is based on either grade 3 achievement or grade 4 teachers respectively. However, if neither of those conditions hold, then the Rothstein test no longer rejects, as can be seen in column 6. Thus, under these conditions the Rothstein test does suggest that either double-lagged achievement or lagged teachers were left out of the VAM. While this might seem encouraging for the use of the Rothstein test, the conditions are not plausible because in order for the grade 3 and 4 achievement levels to be linearly related, we had to set the grade 3 and 4 teacher effects to 0 (Appendix B explains why this works). If grade 3 and 4 teachers do affect achievement, then the relationship between the grade 3 and 4 achievement levels is no longer linear. Under these more plausible conditions the Rothstein test will reject even if double lagged achievement and lagged teachers did not affect tracking decisions. This is illustrated by the results in Column 3. To summarize, given a more reasonable data generating process, the Rothstein test cannot be used to identify the potential that key explanatory variables, correlated with tracking, have been left out of the VAM.

In columns 7, 8, and 9 (under "Negatively Correlated Errors"), we present conditions under which the Rothstein test did identify bias. In column 7, we generated data with no tracking. Therefore there was no bias and the Rothstein test did not reject. This is in spite of the fact that the model had negatively correlated errors and lagged teacher effects so that the grade 3 and 4 test scores were no longer linearly related. When there was tracking (as in column 8), there was bias and the Rothstein test appropriately rejected. However, the amount of bias may have little policy relevance as the correlation between the estimated and true teacher effects was close to 1 after adjusting for estimation error.[38] In addition, we did not find statistically significant bias here though this was due only to a lack of precision.[39]

The small amount of bias found in column 8 is important because the data generation process used for this model is quite similar to what was reported in Rothstein (2010) and consistent with other literature.

- The standard deviation of the error is 0.4, as used by Rothstein in his baseline model in Appendix C of his 2010 paper.
- The serial correlation between contiguous errors is around -0.25, which is in the range of what Rothstein finds and is what he used in his baseline model.
- The correlation between errors across two periods is 0, as Rothstein finds.
- The standard deviation of teacher effects is 0.1. This is the value Rothstein used for his baseline model and is in the range of the estimates reported by Hanushek and Rivkin (2010).

The results in column 9 help to illustrate why the model used in column 8 resulted in so little bias. In particular, in column 9, we show that if the baseline scores and double-lagged scores were linearly related (that is, there were no lagged teacher effects), then there would be no bias. Under those conditions, the baseline test can control for the negatively correlated errors (as discussed in Appendix B). While it is not likely that baseline and double-lagged scores are linearly related, they may be almost linearly related given the small magnitude of the lag teacher effects relative to the overall variance of achievement.

If we knew that the error terms were negatively correlated but were unsure if there was tracking, then the Rothstein test would provide evidence of at least some bias. However, *given that tracking in schools is quite likely, evidence of negatively correlated errors is itself evidence of bias*. And, as noted earlier, one cannot use the Rothstein test to check for bias caused by negatively correlated errors because it will reject even in their absence, as illustrated by the results in column 3 of Table 1.

Condition 3 (illustrated by the results in columns 3 and 8 of Table 1) is potentially the most serious shortcoming of the Rothstein test because it means that we cannot use the Rothstein test to identify variables left out of a VAM. Thus, it is reasonable to ask whether Condition 3 will cause the Rothstein test to reject under different parameterizations of the model. After all, if, under plausible scenarios, the non-linearity will not cause Rothstein to inappropriately falsify then Rosthein's test falsifying a VAM may well be indicative of variables that might cause true bias. We considered the following variations. First, we added measurement error. The results remained the same in the sense

[38] We also ran models with much larger teacher effects (standard deviation of 0.75), larger negative correlations (-0.90), and both. The adjusted correlations between the estimated and true teacher effects remained at 0.93 and above in these models.

[39] With larger sample sizes we do find bias. Also in results available upon request we simulated data for a model very similar to the "baseline" model in Appendix C of the Rothstein paper. We find biased estimates for this model which has negatively correlated errors and lag teacher effects. The correlations between the estimated and true teacher effect estimates remain well over 0.90 after adjusting for estimation error. This model includes 1.2 million teachers and measurement error, as well as teacher and school effects that are correlated across grades.

that the Rothstein test continued to reject and the standard deviation of the estimated future teacher effects actually increased in size.[40] Second, we added three polynomial terms to the model with measurement error.[41] In this case the Rothstein 2 test did not reject. However, when we increase the sample size of students per teacher from 20 to 60 it does reject. Also, the other versions of the Rothstein test rejected in all models estimated and the standard deviation of the estimated future teacher effects remained larger than the standard deviation of the true teacher effects for the Rothstein 1 test and at least half as large for the other two. Third, we reduced the amount of tracking by increasing the standard deviation of the random variable that affects tracking from 0.2 to 1.0. In this case the Rothstein 1 test (the one presented in Table 1) did not reject, but when we increased the number of students per teacher to 60 it does. In addition, the estimated future teacher effects remained about the same when we varied the amount of tracking.[42] Lastly, we ran a model with measurement error and the reduced level of tracking but with 200 students per teacher to get a better sense of the magnitude of the future teacher effects when there is less estimation error. The standard deviations of the future teacher effects remained about half as large as the standard deviations of the true teacher effects for the Rothstein 1 and 2 tests.[43] These magnitudes are similar to those reported by Rothstein in his paper, after adjusting for estimation error.[44] To summarize, there are conditions under which the Rothstein test will not reject, but in most of the cases we identified, the rejection returns when we increase the sample size and the magnitude of the standard deviation of the estimated future teacher effects remains at least about half as large as the standard deviation of the estimated true teacher effects.

In Table IV of his paper Rothstein presents results for the Rothstein 1 and 2 tests discussed here. Rothstein estimated his models with 2,700 teachers compared to only 800 in our models. Rothstein reports p-values ranging from 0.001 to 0.162 depending on the model estimated. In comparison all of the results we report as being statistically significant have p-values of 0.05 level or below. In addition, he reports standard deviations of the estimated future teacher effects of between 0.120 and 0.150, all in the range of values we find. The bottom line is that the Rothstein results appear to be similar to what we find based on Condition 3 which suggests that they may be driven by a similar type of non-linearity.

In his paper Rothstein also notes that his test can fail to falsify when there is bias. The results in columns 10 and 11 (under "Failing to Falsify") support this point. We obtained these results by creating an omitted variable that affected both grade 5 achievement and the selection of the grade 5 teachers, but was uncorrelated with lagged achievement and lagged teachers. Because the variable

[40] The other versions of the Rothstein test also continued to reject and the standard deviations of the future teacher effects increased compared to the model in Column 8 of Table 1. The measurement error had a standard deviation of 0.2, as used by Rothstein in his baseline model. We adjusted the variance of the grade 1 achievement and coefficient on lagged achievement so that the achievement scores remained with standard deviations close to 1.

[41] We added variables as controls equal to the achievement variable on the right hand side, squared, cubed, and to the power of 4.

[42] We estimated 14 models varying the standard deviation of the tracking error from 0.2 to 2.0. The standard deviation of the future teacher effects for the Rothstein 1 remained above 0.12 for all models estimated.

[43] They were substantially larger for the Rothstein 3 test. The estimates were statistically significant for the Rothstein 1 and 3 tests but not for Rothstein 2.

[44] He found ratios of the standard deviation of future teacher effects to true teacher effects, adjusting for estimation error, that ranged from 0.45 to 0.75.

was uncorrelated with lagged achievement, it did not cause the Rothstein test to reject. Column 10 presents results with uncorrelated errors. In column 11, we show estimates for a model with negatively correlated errors that also did not cause the Rothstein test to reject.

Random assignment of *individual* students to teachers is a clear case in which VAMs yield unbiased estimated teacher effects, and the Rothstein test appropriately fails to falsify. Interestingly, however, random assignment of *groups* of students to teachers can cause the Rothstein test to falsify if students were tracked into those groups. This is shown in the last column of Table 1, which reports findings generated when students were tracked into classrooms based on their previous classrooms alone, but teacher assignment to classrooms was random. The same arguments for the Rothstein test rejecting hold as for the previous models. Indeed, the first three columns of Table 1 are relevant because they simply describe how students were tracked into grade 5 classrooms, but not how grade 5 teachers were assigned to those classrooms.[45] Thus, they would hold if teachers were randomly assigned.

IV. CONCLUSION

As we noted in the outset of this paper, Rothstein's critique of value-added methods used to estimate teacher effectiveness has been cited by both research and policymaking communities as a reason to doubt the wisdom of using VAMs for high-stakes purposes. The findings we present here, however, call into question whether the Rothstein falsification approach provides accurate guidance regarding the magnitude or existence of bias of teacher effect estimates.

Ideally, the Rothstein test could be used to identify VAMs that produce biased estimates of current teacher effects. We do show that Rothstein's test might be useful for identifying important control variables left out of a VAM, but find that it only does this under conditions that are not plausible. More precisely, we find that one cannot use the Rothstein test to reject the hypothesis that students were effectively randomly assigned conditional on lagged achievement. In addition, we find that when data are generated that appear similar to the data analyzed by Rothstein, estimated future teacher effects, from his tests, are similar in magnitude to the true teacher effects, but the bias for current teacher effects is extremely small, suggesting that the magnitude of the future teacher effects does not provide useful information about the magnitude of the bias. In a nutshell, the Rothstein test can be used to identify the existence of tracking, but the tracking could well be a function of lagged achievement, the variable that is included in most VAMs. It does not appear that the Rothstein test can be used to tell us much more.

We would argue that Rothstein's 2010 paper raised important concerns about the ability of VAMs to produce unbiased estimates of teacher effectiveness, but the Rothstein test *itself* does not provide useful guidance regarding VAMs. Given this, we believe that more work needs to be done to understand the potential reasons why VAMs might produce biased teacher effect estimates. This will likely involve a closer look at the various factors affecting student sorting into classrooms so

[45] In results available upon request, we ran simulations that align with each of those presented in Table 1, but using 50,000 students per teacher and only two teachers and one school each. The results were generally similar to those in Table 1. An important exception is that we did find bias for column 8 as expected.

that one can better account for student sorting when estimating teacher effects.[46] In fact, there is a spate of recent work that touches on this issue (Feng 2010; Gao 2011; Guarino et al. 2011; Jacob and Lefgren 2007; Kraemer et al. 2011). Simulation evidence also shows that, properly specified, VAMs produce estimates of teacher effectiveness that are close to true values under a range of sorting mechanisms.

From a policy perspective, the important question may not be whether there is any bias, but the potential magnitude of any bias. It is quite likely that teacher effectiveness estimates generated from VAMs are biased to some degree but, as shown in Rothstein (2009), Kinsler (2011), and our simulations, the magnitude of bias may be relatively inconsequential. Decisions about using VAMs should consider how this bias compares to potential information that value-added models can provide about teacher effectiveness over, or in addition to, other means of assessment.[47]

[46] Ashenfelter (1978) makes a similar point in his paper, which looks at similar issues outside of value-added models.

[47] Value-added estimates may also be very imprecise (Schochet and Chiang 2010). Other measures of teacher effectiveness may also be imprecise, so the policy focus should probably be on how best to obtain more precise estimates of teacher performance. This may involve using some combination of VAM and non-VAM measures. Indeed, if well implemented such combinations may be optimal both for reducing bias and for improving precision.

REFERENCES

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, vol. 60, 1978, pp. 47–50.

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. "Problems with the Use of Student Test Scores to Evaluate Teachers." Briefing Paper #278. Washington, DC: Economic Policy Institute, August 29, 2010.

Ballou, Dale, William L. Sanders, and Paul S. Wright. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Education and Behavioral Statistics*, vol. 29, no. 1, spring 2004, pp. 37–65.

Briggs, D., and B. Domingue. "Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the *Los Angeles Times*." Boulder, CO: National Education Policy Center, 2011. Available at http://nepc.colorado.edu/publication/due-diligence. Accessed October 11, 2011.

Chaplin, Duncan, Shinu Verghese, Hanley Chiang, Kathy Sonnenfeld, Margaret Sullivan, Barbara Kennen, Virginia Knechtel, John Hall, and Dominic Harris. "2008 Principal/Vice Principal Survey: Results for Evaluation of the Effective Practice Incentive Community (EPIC) Final Report." Washington, DC: Mathematica Policy Research, March 30, 2009.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "The Long-term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Working Paper No. 17699. Cambridge, MA: National Bureau of Economic Research, December, 2011.

Eckert, Jonathan M., and Joan Dabrowski. "Should Value-Added Be Used for Performance Pay?" *Phi Delta Kappan*, vol. 91, no. 8, May 2010, pp. 88–92.

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

Fuller, W.A. *Measurement Error Models.* New York: John Wiley & Sons, 1987.

Gao, Niu. "School Accountability, Principal Characteristics and Teacher Assignment," Unpublished manuscript, Florida State University. 2011.

Goldberger, Arthur J. *A Course in Econometrics.* Cambridge, MA: Harvard University Press, 1991.

Goldhaber, Dan, and Michael Hansen. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review*, vol. 100, no. 2, 2010a, pp. 250-255.

Goldhaber, Dan, and Michael Hansen. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Center for Education Data and Research Working Paper #2010-3. Seattle, WA: University of Washington, 2010b.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. *Identifying Effective Teachers Using Performance on the Job*. Washington, DC: The Brookings Institute, 2006.

Guarino, C., Reckase, M., & Wooldridge, J. "Can Value-added Measures of Teacher Performance be Trusted?" Unpublished manuscript. 2011.

Haertel, Edward, Jesse Rothstein, Audrey Amrein-Beardsley, and Linda Darling-Hammond. "Getting Teacher Evaluation Right: A Challenge for Policy Makers." Capitol Hill Briefing, September 14, 2011. American Education Research Association and National Academy of Education. Available at [http://edpolicy.stanford.edu/publications/pubs/421 and http://www.aera.net/uploadedFiles/Gov_Relations/AERA-NAE_briefing_Combined_Slides_FOR_PRINTING.pdf]. Accessed October 29th 2011.

Hanushek, Eric A. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press, 2009.

Hanushek, Eric A, and Steven G. Rivkin. "Using Value-Added Measures of Teacher Quality." National Center for Analysis of Longitudinal Data in Education Research (CALDER) Brief 9. Washington, DC: Urban Institute, May 2010.

Harris, Douglas, and Tim Sass. "Teacher Training, Teacher Quality, and Student Achievement." Working paper. Tallahassee, FL: Florida State University, 2010.

Harris, Douglas, Tim Sass, and Anastasia Semykina. "Value-Added Models and the Measurement of Teacher Productivity." National Center for Analysis of Longitudinal Data in Education Research (CALDER) Working Paper No. 54. Washington, DC: Urban Institute, 2010.

Heckman, James, and Joseph Hotz. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impacts of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association,* vol. 84, issue 408, 1989, pp. 862-874.

Jacob, Brian A., and Lars Lefgren. "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers." *Quarterly Journal of Economics,* vol. 122, no. 4, 2007, pp. 1603–1637.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kinsler, Josh. "Assessing Rothstein's Critique of Teacher Value-Added Models." Working paper. February 2011.

Koedel, Cory, and Julian R. Betts. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." University of Missouri Working Paper 09-02. Columbia, MO: University of Missouri, 2009.

Kraemer, Sara, Robin Worth, and Robert H. Meyer. "Classroom Assignment Practices in Urban School Districts Using Teacher Level Value-Added Systems." Working paper. Madison, WI: Value-Added Research Center, University of Wisconsin-Madison, 2011.

McCaffrey, Daniel F., Daniel Koretz, J.R. Lockwood, Thomas A. Louis, and Laura S. Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics,* vol. 29, no. 1, March 20, 2004, pp. 67–101.

McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood, and Kata Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, October 14, 2009, pp. 572–606.

Meyer, Robert H. "The Production of Mathematics Skills in High School: What Works?" In *Earning and Learning: How Schools Matter,* edited by S. Mayer and P. Peterson. Washington, DC: The Brookings Institution, 1999.

Podgursky, Michael, and Mathew Springer. "Teacher Performance and Pay: A Review." *Journal of Policy Analysis and Management*, vol. 26, no. 4, 2007, pp. 909–949.Potamites, Liz, Kevin Booker, Duncan Chaplin, and Eric Isenberg. "Measuring School and Teacher Effectiveness in the EPIC Charter School Consortium—Year 2 Final Report." Washington, DC: Mathematica Policy Research, October 2009.

Rothman, Robert. "Beyond Test Scores: Adding Value to Assessment." *The School Administrator,* vol. 67, no. 2, February 2010, pp. 20–24.

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics,* vol. 125, no. 1, February 2010, pp. 175–214.

Rothstein, Jesse. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy*, vol. 4, no. 3, 2009, pp. 537–571.

Schochet, Peter Z., and Hanley S. Chiang. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." NCEE 2010-4004. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

Spearman, C. "The Proof and Measurement of Association Between Two Things." *The American Journal of Psychology*, vol. 15, no. 1, 1904, pp. 72–101.

Theil, H. "Specification Errors and the Estimation of Economic Relationships." *Review of the International Statistical Institute,* 1957, pp. 41–51.

Todd, Petra E., and Kenneth I. Wolpin. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal,* vol. 113, no. 485, February 2003, pp. F3–F33.

Viadero, Debra. "Princeton Study Takes Aim at 'Value-Added' Measure." *Education Week*, vol. 17, no. 11, June 23, 2009, pp. 6–11.

APPENDIX A

WHEN FALSIFICATION TESTS FAIL

In this appendix, we elaborate in more detail why the Rothstein test can incorrectly falsify correctly specified VAMs. As discussed in the main body of this paper, the omission of variables used for tracking students need not cause bias if they do not impact current test scores directly, although the existence of such omitted variables in combination with negatively correlated error terms would suggest bias. Conditions 1 and 2 suggest that the Rothstein test could be used to help identify the existence of such omitted variables—in particular, double-lagged achievement and lagged teachers. Here, however, we illustrate that the Rothstein test cannot be used to rule out the possibility that students were tracked randomly conditional on lagged achievement. To do this, we use a one-school, two-teacher example in which tracking decisions depend only on lagged achievement and a random error, and there are no omitted variables from the model. We start with an equation for R_{1g}, the coefficient on the better (that is, more effective) "future" teacher from the Rothstein test (see footnote 22).

$$R_{1g}= \mathrm{cov}(\tau_{1ig},A_{i(g-2)},\ |A_{i(g-1)})/\mathrm{var}(\tau 1_{ig}|A_{i(g-1)}) = \mathrm{cov}(\tau^*_{1ig},A^*_{i(g-2)})/\mathrm{Var}(\tau^*_{1ig})$$

where $A^*_{i(g-2)}$ and τ^*_{1ig} are the residuals that result from regressing $A_{i(g-2)}$ and τ_{1ig} on $A_{i(g-1)}$.

As shown in the main text, one component of $\mathrm{cov}(\tau^*_{1ig},A^*_{i(g-2)})$ is $\mathrm{cov}(\tau^*_{1ig},e^*_{i(g-1)})$. If this component is non-zero, then the Rothstein test can reject incorrectly (condition 3).

Condition 3 would not occur if lagged achievement and its error term were jointly normally distributed. In this situation, the expected value of the error would be a linear function of lagged achievement and the residual error would be uncorrelated with any function of lagged achievement (linear or nonlinear).[48] But, as we show below, it is quite likely that lagged achievement is not normally distributed because it is itself impacted by lagged teachers. The impacts of these teachers depend on the fraction of time a student is assigned to a teacher, which is necessarily bounded between zero and one, and therefore is not normally distributed. This means that the expected value of the lagged error (one element in the equation for condition 3) is likely to be a nonlinear function of lagged achievement. The current teacher (the other element of condition 3) is also a nonlinear function of lagged achievement because assignment to the current teacher is also a non-normally distributed variable, that is, students either are, or are not, assigned to a given teacher. Since both the lagged error and current teacher are nonlinear functions of lagged achievement, they can remain correlated even after conditioning on lagged achievement in a linear regression.

In showing a more formal description of how Rothstein's test can reject because of condition 3, we make the simplifying assumption that the lagged error is jointly normally distributed with double

[48] If the current teacher is a function of lagged achievement and an uncorrelated error, its residual (controlling for lagged achievement) would also be uncorrelated with the lagged error.

lagged achievement.[49] Given this joint normality assumption, we know that for either lagged teacher's students, the expected value of $A_{i(g-2)}$ is a linear function of $A_{i(g-1)}$. Let these functions be:

$E(A_{i(g-2)} | \tau_{1i(g-1)} = 1, A_{i(g-1)}) = \alpha_1+\lambda_1 A_{i(g-1)}$ for students with the better lagged teacher, and

$E(A_{i(g-2)} | \tau_{1i(g-1)} = 0, A_{i(g-1)}) = \alpha_0+\lambda_0 A_{i(g-1)}$ for students with the omitted lagged teacher.

Now consider the function for the probability of having the more effective lagged teacher as a function of lagged achievement. This is nonlinear since $\tau_{1i(g-1)}$ is a discrete variable. Thus,

$$\tau_{1i(g-1)} = T_1(A_{i(g-1)})$$

Note that this is not a tracking function because we are describing a function that relates the lagged teacher to the lagged test score—and not to the double-lagged test score.

Finally, the equation for the expected value of $A_{i(g-2)}$ as a function of $A_{i(g-1)}$ that combines both sets of students can be written as follows:

$$E(A_{i(g-2)} | A_{i(g-1)}) = (\alpha_1+\lambda_1 A_{i(g-1)})T_1(A_{i(g-1)}) + (\alpha o+\lambda_0)A_{i(g-1)}T_0(A_{i(g-1)})$$

where $T_0(A_{i(g-1)}) = 1- T_1(A_{i(g-1)})$.

$T_0()$ and $T_1()$ are both nonlinear functions of $A_{i(g-1)}$. Thus, $E(A_{i(g-2)}|A_{i(g-1)})$ is a nonlinear function of $A_{i(g-1)}$. Since both $E(A_{i(g-2)}|A_{i(g-1)})$ and τ_{ig} are nonlinear functions of $A_{i(g-1)}$, this suggests that they could be correlated even after controlling for $A_{i(g-1)}$ linearly. This can, in turn, cause the Rothstein test to reject incorrectly.[50]

[49] We use this example to show how condition 3 might hold because neither of the variables that appear in conditions 1 or 2 affect tracking in this example (double-lagged achievement or the lagged teacher). However, it is also true that conditions 1 and 2 might hold in this situation. This can happen because of nonlinearities similar to those discussed here.

[50] Allowing $A_{i(g-2)}$ to be non-normal could introduce additional nonlinearities that could also cause the Rothstein test to reject incorrectly.

APPENDIX B

NEGATIVELY CORRELATED ERRORS NEED NOT CAUSE BIAS

In this appendix, we show that negatively correlated errors need not cause bias if lagged achievement scores are normally distributed. To make this point, we first show that negatively correlated errors may result in an omitted variable that is a linear function of the lagged achievement score. An omitted variable with this property will bias the lagged achievement coefficient estimate, but will not cause bias in the teacher effects, a point that is well known. This is important because it means that the negative correlation in errors across grades assumed by Rothstein need not cause bias on its own.

To investigate these issues, we consider a model in which the *error terms are negatively correlated*, as Rothstein posits. This could happen if students who had an above-average error term in the previous period forget more than students with an average error term in direct proportion to how far above average they were in the previous period. Similarly, those who had a below-average error term in the previous period learn more than students with an average error term (perhaps from other students or their teacher), again, in direct proportion to how far they were below average in the previous period.[51] Mathematically, this can be written as:

$$e_{ig} = \omega_1 u_{ig} + \omega_2 u_{i(g-1)}$$

where $\mathrm{cov}(u_{ig}, u_{i(g-1)}) = 0$ and $\omega_2 < 0$.

This implies that

$$\text{(B.1)} \qquad \mathrm{cov}(e_{ig}, e_{i(g-1)}) = \omega_2 \mathrm{var}(u_{i(g-1)}) < 0.$$

As in our simulations, we assume that lagged achievement depends only on lagged teachers, an error term, and a random starting point ($A_{i(g-2)}$) that is normally distributed.[52] Thus:

$$A_{ig} = A_{i(g-1)}\lambda + \tau_{1ig}\beta_{1g} + e_{ig}$$

$$A_{i(g-1)} = A_{i(g-2)}\lambda + \tau_{1i(g-1)}\beta_{1(g-1)} + e_{i(g-1)}$$

To generate normally distributed baseline scores (and unbiased estimated teacher effect estimates), we assume that lagged teachers have no impact on lagged achievement levels. Thus,

$$A_{i(g-1)} = A_{i(g-2)}\lambda + e_{i(g-1)}$$

where $A_{i(g-2)}$, $e_{i(g-1)}$ and e_{ig} are jointly normal and uncorrelated with each other by assumption. This implies that $A_{i(g-1)}$ and e_{ig} are also jointly normal and linearly related because a linear function of two

[51] This is sometimes described as "regression to the mean," although that phrase is sometimes used to describe situations in which the true errors are uncorrelated across grades.

[52] This can be justified if we think of two grade levels in the past as the grade when the child entered school and that there was no tracking in that grade. All subsequent learning is captured by later teacher effects and error terms.

jointly normally distributed variables is also jointly normal and linearly associated with each of those variables (Goldberger 1991; Theil 1957). Indeed, all four variables are jointly normal and linearly related.

$$\begin{pmatrix} A_{i(g-1)} \\ A_{i(g-2)} \\ e_{i\,g} \\ e_{i(g\,-\,1)} \end{pmatrix} \sim \mathrm{N}(\mu, \Sigma)$$

In particular, the expected value of e_{ig} is a linear function of $A_{i(g-1)}$. Given this, let er_{ig} be the residual from a regression of e_{ig} on $A_{i(g-1).}$

$$e_{ig} = \gamma A_{i(g-1)} + er_{ig}$$

where γ is the coefficient on lagged achievement.[53]

Now we need to show that er_{ig} is uncorrelated with τ_{ig}, the current teacher. We start by assuming a specific functional form for $\tau_{tig,}$

(B.2) $\tau_{1,i,g} = 1$ if $A_{(i,g-1)} > 0$ and 0 otherwise.

We then show that er_{ig} is uncorrelated with any function of $A_{i(g-1)}$ and therefore is uncorrelated with $\tau_{1,i,g}$.

By assumption, e_{ig} and $A_{i(g-1)}$ are jointly normal. By construction, er_{ig} is a linear function of these variables equal to e_{ig}-λA_{ig}. Thus er_{ig} and $A_{i(g-1)}$ are also jointly normal. By construction, er_{ig} is also uncorrelated with $A_{i(g-1)}$. Joint normality and zero correlation implies independence. This, in turn, means that er_{ig} is uncorrelated with any function of $A_{i(g-1)}$ regardless of whether it is linear or nonlinear. Since $\tau_{tig,}$ is a function of $A_{i(g-1)}$, it is also uncorrelated with er_{ig}.[54]

Using the symbols from equation 2 in the main body of this paper (the omitted variable bias formula), this means that cov(e^*_{ig},$\tau^*_{ti,g}$) = 0. To see this, note that er_{ig} is the residual that remains after regressing e_{ig} on $A_{i(g-1)}$.[55] Thus, er_{ig} is the same as e^*_{ig}. Similarly, $\tau^*_{1i,g}$ is the residual that remains after regressing $\tau_{1i,g}$ on $A_{i(g-1).}$ If e^*_{ig}, is uncorrelated with $\tau_{1i,g}$ then it will also be uncorrelated with $\tau^*_{1i,g}$ because τ^*_{1ig} is a linear function of τ_{1ig} and $A_{i(g-1)}$ and e^*_{ig} is uncorrelated with both of those variables.

[53] We expect γ to be less than 0 since cov(e_{ig},$e_{i(g-1)}$) is less than 0.

[54] In a more realistic scenario, τ_{1ig} would also depend on some additional variables. As long as they are also distributed independently of er_{ig}, then er_{ig} will remain conditionally uncorrelated with τ_{ig}.

[55] In the main body of this paper, we discussed creating residuals by regressing each variable on lagged achievement and the other teacher dummies. In this model, there are no other teachers because there are only two teachers and one is omitted.

The negative correlation in errors means that students who scored lowest on the previous test will score somewhat higher than otherwise expected in the current period and vice versa. The coefficient estimate on $A_{i(g-1)}$ will be biased downwards, but, in this case, *the negative correlation has no impact on the coefficient on τ_{1ig}.*

Some readers might also be concerned about the plausibility of the data generation process we propose for tracking because it appears to depend on a latent variable that is a linear function of lagged achievement. However, as noted above, tracking is necessarily a nonlinear function of this latent variable. The functional form we propose allows for this. More precisely, one could think of tracking as a two-stage system in which tracking depends on a latent variable that, in turn, depends on lagged achievement. This can be written either by stage or in a single stage by substituting out for the latent variable (LV). Thus,

Stage 1: $LV = \beta T^{*}A_{i(g-1)}$

Stage 2: $\tau_{1ig} = T(LV)$

Combined: $\tau_{1ig} = T(\beta T^{*}A_{i(g-1)})$

where βT^{*} is the coefficient on lagged achievement in Stage 1.

We have assumed that the first stage is linear. However, even if the first stage were nonlinear this would not affect our argument because the second stage is nonlinear. Thus, nonlinearity in the first stage is not an issue. What is key to our argument for this appendix—that a "plausible" data generation process can yield unbiased results—is that the relationship between the current error term and lagged achievement is linear.

APPENDIX C

RESULTS FOR ALL THREE VERSIONS OF THE ROTHSTEIN TEST

In the main body of the paper we presented results for one version of the Rothstein test, which we refer to as the Rothstein 1 test. In this appendix we present results for all three versions of the Rothstein test described in the main body of the paper.

As shown in Table C.1, the results are similar across the three tests in the sense that when one test rejects the others do as well. In addition, the estimated future teacher effects have standard deviations that are similar in magnitude to or larger than the standard deviations of the estimated true teacher effects. There are, however, a few interesting exceptions. First, the Rothstein 3 test has much larger test statistics and future teacher standard deviations than the other two tests and it rejects in Columns 6, 9 and 11 when Rothstein 1 does not. This is likely due to the fact that the Rothstein 3 test rejects for at least one reason that does not hold for the Rothstein 1 or 2 tests. In particular, the lagged test score, which is the outcome for the Rothstein 3 test, is, in theory, the variable that was used to track the current teacher based on a standard VAM. Thus, the current teacher is effectively an endogenous variable in the Rothstein 3 test.[56]

The Rothstein 2 test also rejects in columns 6, 9 and 11 when the Rothstein 1 test did not. This suggests that rejection is occurring for different reasons in the Rothstein 2 test compared to the Rothstein 1 test. This may be related to non-linearities that are created when one adds in the lagged teachers as control variables. In the other direction, the Rothstein 2 test does not reject in column 12 even though Rothstein 1 and 3 do reject. This is likely due to the fact that the Rothstein 2 test effectively controls for classroom tracking which may be the cause of rejection for the Rothstein 1 and 3 tests.

[56] A similar point holds for the falsification test Rothstein uses for the VAM1 model in his paper.

Table C.1. Simulation Results for Rothstein Falsification Tests and Bias Test, by Model and Test Type

Parameter	Results by Condition			Linear Falsification Test			Negatively Correlated Errors			Failing to Falsify		Rejecting RA
Result	1	2	3	4	5	6	7	8	9	10	11	12
Teachers Tracked on	A_4, A_3	A_4,τ_4	A_4	A_4,A_3	A_4,τ_4	A_4	Rand	A_4	A_4	ov$^*_g$	ov$^*_{g,}$,A_4	τ_4
Cor(e_g,e_{g-1})= Cor(e_{g-1},e_{g-2})							-0.25	-0.25	-0.25		-0.25	
Var(ov$^*_g$)										0.2	0.2	
Std Dev(β_3, β_4)	0.1	0.1	0.1				0.1	0.1		0.1		0.1
Bias Test	0.95	0.95	0.98	1.01	0.96	0.92	1.01	1.00	0.88	**3.34***	**2.02***	1.02
Rothstein 1 Test	**8.34***	**8.91***	**1.27***	**7.11***	**7.97***	1.03	0.99	**1.18***	0.96	1.00	0.97	**2.01***
Rothstein 2 Test	**1.19***	**2.08***	**1.98***	**1.12***	**2.06***	**1.97***	1.00	**2.04***	**1.96***	0.99	**1.67***	0.95
Rothstein 3 Test	**7.57***	**3.73***	**22.68***	**7.74***	**2.93***	**20.83***	0.93	**22.56***	**20.6***	0.96	**17.82***	**2.19***
Std Dev Rothstein 1 Estimates	0.331	0.292	0.149	0.311	0.278	0.130	0.129	0.144	0.125	0.133	0.126	0.174
Std Dev Rothstein 2 Estimates	0.168	0.178	0.177	0.167	0.177	0.169	0.131	0.169	0.164	0.134	0.153	0.127
Std Dev Rothstein 3 Estimates	0.325	0.238	0.368	0.320	0.195	0.343	0.120	0.354	0.345	0.127	0.316	0.189

Notes: The three versions of the Rothstein test are described above.

A_g is achievement in grade g, τ_g is a vector of teacher dummies, ov$^*_g$ is an omitted variable that causes bias, e_g is an error term, and β_g is a vector of coefficients on the teacher dummies. Blanks indicate zeros. Tracking is based on the sum of the variables indicated in the table and a random error with standard deviation of 0.2 Rand means this random error was the only one used for tracking. School effects are set to zero. Achievement has a standard deviation close to one. The errors e_g, e_{g-1} and e_{g-2} are in the achievement equations and have standard deviations of 0.4. The variable ov$^*_g$ affects both achievement and tracking in grade 5, but is uncorrelated with previous grade variables. Lambda is 0.91 in all models except for 7, 8, 9, and 11, where it is 0.95. The bias test is a joint test of the difference between the estimated and true teacher effects in a standard VAM. The Rothstein test is described in the text. The test statistics are F-tests. The cut-points for 5 and 1 percent statistical significance given our sample sizes are 1.10 and 1.14 respectively. Test statistics in bold with a "*"are significant at the 5 percent level. The standard deviation of grade 5 teacher effects is 0.1 in all models. Std Dev(β_3, β_4) refers to the standard deviations of the teacher effects in grades 3 and 4. Each model has 200 schools, four teachers per school, and 20 students per teacher. The regressions control for school effects so we only estimate teacher effects for 600 teachers in each model (three teachers per school and 200 schools).

Acknowledgements

The authors wish to acknowledge helpful comments on an earlier draft of this paper from Jesse Rothstein, Mike Hansen, Josh Kinsler, Cory Koedel, Peter Schochet, Tim Sass, Anastasia Semykina, Dale Ballou, and Dick Startz, as well as comments on a draft presentation by Eric Isenberg, Steve Glazerman, Alex Resch, Arif Mamun, Jillian Berk, and Ali Protik. We also would like to thank Stephanie Liddle and Roddy Theobald for their able assistance with this work. The views expressed in this paper are those of the authors and do not necessarily reflect the organizations with which the authors are affiliated. Any errors are solely the authors' responsibility.