ABSTRACT

In a provocative and influential paper, Jesse Rothstein (2010) finds that standard value-added models (VAMs) suggest implausible 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. However, in this paper, we illustrate—theoretically and through simulations—plausible conditions under which the Rothstein falsification test rejects VAMs even when there is no bias in estimated teacher effects, and even when students are randomly assigned conditional on the covariates in the model. On the whole, our findings show that the "Rothstein falsification test" is not definitive in showing bias, which suggests a much more encouraging picture for those wishing to use VAM teacher effect estimates for policy purposes.

I.

Rothstein cited his falsification test results and said that value-added is “not fair to…special needs teachers…[or] other specialists.”[3]

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,

β_t is the impact of teacher t compared to the omitted teacher,[5] Rothstein questions whether standard VAMs produce unbiased estimates of teacher effect on student learning.[9] 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 $A_{i(g-1)}$.[10]

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,

$$e_{ig} = J ov_{ig} + eu_{ig}$$

where J is the coefficient on ov_{ig}. J 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:

$$\text{Bias}(\hat{E}_t) \quad E(\hat{E}_t) \quad E_t \quad J S_{te}$$

where

= estimate of and

S_{te} = coefficient on W_{tig} from a regression of ov_{ig} on all right-hand side variables in equation 1 (except e_{ig}).

It can be shown that,

9

than students who are otherwise similar.[16]

C. Rothstein's Falsification Test

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)[18] $A_{ig} = \text{Œ}A_{i(g-1)} + E_{1g} W_{1ig} + e_{ig}$

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

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

$W_{1ig} = T(A_{i(g-1)})$[19]

Rothstein's falsification test is based on a regression of lagged achievement on current achievement and future teachers.[20]

The Rothstein test using future teachers can 04 Te(e)1(4(r)7(e)1(475.56 Tm[(2)1(0)]TJEMC e)]TJr)3(.5(e)1iT(l)3

The numerator in R_{1g} is the covariance between current teachers and the residual from a regression of double-lagged achievement on current achievement.[22]

$$(5)\quad A_{i(g-2)} = O \; {}_{ig}A + u_{i(g-2)}$$

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

$$\mathrm{cov}(W_{1ig}, u_{i(g-2)} \mid A_{i(g-1)<d[>d[0t.}$$

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.[33]

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. The second set of columns (under "Linear Falsification Test") covers findings when grade 4 achievement was

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

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.[37] In addition, we did not find statistically significant bias here though this was due only to a lack of precision.[38]

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).[39] While it is not likely

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 is not possible to determine whether the test is rejecting because of the negative correlation or because of one of the other conditions specified above.

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 bias of teacher effect estimates.

Ideally,

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-

Haertel, Edward, Jesse Rothstein, Audrey Amrein-Beardsley, and Linda Darling-Hammond. “Getting Te2.53 ob 0.Paginf E.53 H753 onS53 6vi1yPolicy Right: Makers’ Challenge for Capitol Hill Briefing, September 14, 2011. American Education Research Association and National Academy of Education. Aailabl713(e)11(at)6( )]TJ0 Tc 0 Tw ([)TjEMC ET/Span <</MCII

NAiefing Combi753 ned Sli753 des FOR PRIN]TING.753 pdf

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."

lagged achievement.[46]

E(A

To generate normally distributed baseline scores (and unbiased estimated teacher effect estimates),

About the Series

Policymakers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers policymakers and researchers access to our most current work.

For more information, contact Duncan Chaplin, senior researcher, at dchaplin@mathematica-mpr.com.

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.