IRLE
Institute for Research on Labor and Employment

Policy Brief

December 2016

Can value-added models identify teachers' impacts?

Jesse Rothstein

Overview

"Value-added models" (VAMs) purport to be able to identify a teacher's causal effect from data on students' test scores. But with non-random assignment of students to teachers, some teachers may be penalized and others rewarded based on the students that they teach rather than on their own effectiveness. In a recent paper, "Revisiting the impact of teachers," I show that even the most sophisticated VAMs remain importantly biased. I also show that recent conclusions that high-VAM teachers have dramatic effects on students' long-run outcomes are unsupported, as student sorting accounts for much if not all of the teachers' apparent long-run effects. Sorting-adjusted long-run impacts of high-VAM teachers cannot be distinguished from zero. Any policies that use VAM scores for teacher evaluation will need to account for the biases that this introduces, and will have more modest effects – if any – on students' long-run outcomes than has been claimed.

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction

"Value-added models" (VAMs) attempt to identify a teacher's causal effect from data on students' test scores. But teachers' causal effects are identified only if students are as good as randomly assigned to teachers. With non-random assignment of students to teachers, some teachers may be penalized and others rewarded based on the students that they teach rather than on their own effectiveness.

A recent study of millions of students in New York City claims to settle the question.[1] The authors, Raj Chetty, John Friedman, and Jonah Rockoff, introduce a quasi-experimental strategy that they argue demonstrates that teachers' VAM scores are essentially unbiased by student sorting. A companion paper by the same authors[2] concludes that teachers have large impacts on students' long-run outcomes, such as completed education and adult earnings, and that a teacher's value-added is a strong proxy for these impacts. All of this points to an important role for VAM scores in teacher compensation and retention decisions.

In a recent paper, "Revisiting the impact of teachers,"[3] I show that the quasi-experimental estimates are not decisive, and if anything point toward the presence, not the absence, of important bias in VAM scores. The quasi-experimental strategy relies on an assumption that teachers switch schools and grades at random, but this is demonstrably incorrect. When adjusted for this, quasi-experimental estimates indicate that VAM scores are indeed biased by student sorting. Moreover, this sorting also accounts for much if not all of the apparent effect of high-VAM teachers on students' long-run outcomes. Sorting-adjusted impacts cannot be distinguished from zero. Any policies that use VAM scores for teacher evaluation will need to account for the biases that this introduces, and will have more modest effects – if any – on students' long-run outcomes than has been claimed.

1. What is "value-added" and why does it matter?

Value-added models (VAMs) are statistical models that attempt to distinguish a teacher's causal impact on her students' learning from other factors – *e.g.*, student ability and out-of-school determinants – that may also influence the students' end-of-year scores. If a teacher's students perform better than predicted based on their prior scores and other characteristics, the teacher is given a high value-added score; if they perform worse than expected, her score is low.

VAMs have been around since the 1970s, but have gained prominence in recent years as improved databases and computing power have made it more feasible to estimate the required models. They have been promoted as objective alternatives or supplements to traditional, more subjective measures of teacher effectiveness such as classroom observations and principal assessments.

Relying on VAM evidence, some scholars have argued that more aggressive teacher quality policies would bring enormous benefits, and that better teachers would translate into dramatic improvements in students' eventual college enrollment, teenage pregnancy, and adult earnings.[4] Eric Hanushek argues that a teacher performing at the 84th percentile is worth $360,000 more than an average teacher per year, and that "US achievement could reach that in Canada and Finland if we replaced with average teachers the least effective 5 to 7 percent of teachers...Accumulated over the lifetime of somebody born today, this improvement in achievement would amount to an increase in total US economic output of $112 trillion in present value."[5]

Many states and school districts now incorporate value-added as a component of teacher evaluations. In some places, high-stakes decisions such as compensation, tenure, and retention are based in large part on teachers' value-added scores.

2. Bias in value-added models

What does bias mean here? Where does it come from?

The use of value-added scores as measures of teacher effectiveness is highly controversial. One of the central concerns – though by no means the only one[6] – is that teachers' value-added scores may be biased by factors outside of their control. If so, value-added-based teacher evaluations will mis-identify effective teachers, diluting their potential impact and forcing teachers to devote their efforts toward gaming the evaluation system rather than focusing on good teaching.

VAM scores can be seen as a comparison of the average performance of a teacher's students on end-of-year tests to predictions of the students' performance with an average teacher. The quality of these predictions is thus of central importance. They are typically based on the student's score on the prior year's test and sometimes in earlier grades, and on a short list of student demographic characteristics (*e.g.*, age, gender, race, and free or reduced price lunch status).

Given these limited variables, there is important variation in student preparedness that is not captured by the prediction model. Some students – those with involved parents, those who were sick on test day on last year, etc. – are quite likely to outperform their predictions, while others – *e.g.*, those with learning disabilities not captured in the VAM predictors – are likely to do less well than predicted. A teacher who is typically assigned students of the former type will receive a value-added score that is biased upward, one that likely overstates her true effectiveness; a teacher who specializes in students who can be expected to underperform their predictions will receive a downward-biased value-added score.

Thus, the presence and magnitude of bias in VAM scores depends on the way that students are assigned to teachers. If students were randomly assigned, VAM scores would be unbiased: Some teachers would receive scores above their effectiveness and others below, but there would be nothing systematic about this. It is very unusual, however, for student-teacher assignments to be random. These assignments in many schools are complex processes, incorporating parental requests, teacher specializations, and assessments of students' individual needs and social dynamics. Some teachers, for example, may be known as especially successful with children who have trouble focusing; others, with delayed readers. The statistical models that are used to construct value-added scores cannot capture such variation and thus these teachers' VAM scores will be biased.

Insofar as such biases are important, they create obvious problems for VAM-based teacher evaluations. A teacher whose VAM score is biased upward due to her student assignments might thereby qualify for bonuses that she has not earned, while one who specializes in tougher students risks being undeservedly dismissed for poor performance. This creates obvious incentives for teachers to avoid the latter assignments in favor of the former, potentially making it harder to staff certain courses and reducing the overall efficacy of the school.

Experimental and quasi-experimental estimates of bias

Researchers are unanimous that classroom assignments are not in general random, and therefore that the conditions that could generate bias are present in most schools. But they have struggled to quantify the magnitude of any resulting biases. I demonstrated their potential importance in a 2010 paper[7], where I showed that 5th grade teachers have substantial apparent "effects" on students' 3rd and 4th grade scores, even after controlling for the factors used to generate VAM predictions. Since these effects cannot be causal, they suggest that the VAM models fail to capture all of the important determinants of classroom assignments. But that strategy did not allow me to precisely quantify the magnitude of the resulting biases, and the results were consistent with either small or large biases in some VAMs. (Other VAMs,

including those in use in many states and districts, were shown to be quite inconsistent with actual assignment processes, with large biases under the best of circumstances.)

Several recent studies have attempted to estimate biases using random assignment experiments.[8] These studies are based on the idea that a VAM score can be used to forecast a teacher's impact in a subsequent year. If the VAM score is unbiased, this should be accurate, on average, so should be a good prediction of the teacher's impact when students are randomly assigned. But if the VAM score is biased, the teacher's impact under random assignment will, under certain assumptions, tend to be closer to average than the forecast. Unfortunately, it has proven very difficult to conduct a random-assignment experiment on sufficient scale, as principals resist requests to assign students to teachers at random. These studies have thus not been able to estimate the magnitude of VAM biases with any precision.

A groundbreaking 2014 study by Raj Chetty, John Friedman, and Jonah Rockoff (hereafter, "CFR") offered a solution to this problem. This study was based on the idea that value-added scores can be used to predict the change in the average performance of the students at a school when a teacher enters or exits the school, and that the accuracy of this prediction can be assessed using data from a large school system without the need to conduct random assignment.

Using data from New York City students over more than a decade, CFR[9] were able to observe many thousands of teacher switches, and thus to obtain very precise estimates. They found no sign that VAM-based forecasts were inaccurate, and concluded that any biases in teacher VAM scores from student sorting are small enough to ignore. In a companion study,[10] CFR also found that the VAM scores of a student's elementary school teachers were correlated with that student's adult outcomes – his or her eventual college attendance, earnings, and even teen pregnancy. From this, they concluded that the stakes in teacher evaluations are very high – that replacing a low-value-added teacher with a high-value-added teacher would lead to dramatic improvements in students' life chances that dwarf the value of the teachers' salaries.

In "Revisiting the Impact of Teachers", I reproduce CFR's analysis in data from North Carolina. I obtain their same results that VAM-based forecasts of the change in student achievement following teacher switching are accurate, on average, and that VAM scores are strongly associated with student's long-run outcomes.

The CFR strategy requires an assumption in place of explicit randomization: Teacher switches between schools, or between grades within schools, must be random – a high-value-added might switch out of a school or grade and be replaced by a low-value-added teacher, or vice versa, and this must be uncorrelated with other determinants of the change in performance of

students in that school and grade. I extend CFR's analysis by carefully testing this assumption. I ask a simple question: Are changes in teacher value-added generated by teacher switching associated with changes in student preparedness, as measured by their test scores in prior grades? In a sample constructed exactly according to CFR's specifications, I find that it does: High-VA teachers tend to replace low-VA teachers when the incoming students have higher prior-grade scores than the previous students, and low-VA teachers tend to replace high-VA teachers when incoming students have lower prior scores than outgoing students.

In a response to my analysis, CFR[11] have suggested that this result is spurious, due to "mechanical" effects related to the use of prior years' student test scores in computing teacher value-added. But this is demonstrably incorrect: The result is robust to a variety of strategies for eliminating such mechanical effects. Moreover, teacher switching is also correlated with changes in student characteristics, such as race and free lunch status, that are not used in the VAM calculation. This result is inconsistent with the mechanical effects explanation.

Together, these results indicate that the teacher switching quasi-experimental strategy is invalid, at least as CFR implement it.[12] Estimates that do not adjust for the association between teacher switching and student preparedness cannot reveal the extent of bias in VAM scores. I consider a variety of methods of implementing such an adjustment. Across all of the approaches I explore, I find evidence of statistically and practically significant bias. Even in the quite sophisticated VAM studied by CFR, teachers' VAM scores are evidently inflated or depressed in part due to the students who they teach, who differ in unobserved ways that are stable over time. This bias accounts for as much as one-third of the variation in teachers' value-added scores, enough to create a great deal of misclassification in VAM-based evaluations of teacher effectiveness.

3. Value-added models and students' long-run outcomes

Perhaps the most important result of CFR's study[13] was that high-value-added teachers have large impacts on their students' long-run outcomes, including educational attainment and earnings. The authors find that, compared to an average elementary or middle school teacher, one who is one standard deviation above average raises each student's earnings by around $350 at age 28. This seems small, but aggregated over the years of the student's career and all of the students that a teacher teaches, it is enormously important.

Like the value-added scores themselves, this is potentially biased by non-random student assignments, and I show that there is no basis for interpreting CFR's long-run impact estimates as causal. Teachers with high value-added scores tend to teach in more advantaged schools and to be assigned students from more advantaged families. Their students' higher earnings may reflect this sorting as much as it does the teachers' impacts on their students' long-run outcomes.

When I re-estimate the long-run effects of high value-added teachers controlling for observable differences in teachers' students I find that they are much smaller, and in quasi-experimental analyses analogous to those used to test for bias – the most reliable strategy – are not distinguishable from zero.

It remains possible that teachers have enormous long-run impacts, and even that these differ systematically between high- and low-value-added teachers. But we simply do not have enough evidence to support this conclusion. Existing strategies are unable to distinguish teachers' long-run impacts from student sorting sufficiently to support a causal interpretation.

Conclusion

In "Revisiting the impact of teachers", I show that recent conclusions regarding the use of teacher value-added models to measure teacher effectiveness are not supported by the evidence. These conclusions were based on a quasi-experiment created by teacher switching, but I show that this switching is decidedly non-random. Adjusting for this leads to quite different conclusions: Teachers' value-added scores are indeed importantly biased by student sorting, and this sorting accounts for much or all of the apparent effect of high-value-added teachers on students' adult outcomes.

This conclusion has an important policy implication: policies that use value-added scores as the basis for personnel decisions are importantly confounded by differences across teachers in the students they teach. Teachers with unusual assignments are at risk of being rewarded or punished for this under value-added-based evaluations, and the use of value-added for high-stakes decision-making will create incentives for teachers to adjust their assignments to their advantage (though to the likely detriment of the school as a whole).

The incorporation of VAMs into teacher evaluations will need to move cautiously, with careful efforts to measure and forestall changes in teacher, principal, and student behavior that would tend to undermine the VAM scores' validity. It will also likely have smaller effects than has sometimes been promised, as raising the average value-added of the entire teacher workforce will not change students' school readiness and thus will have smaller impacts on students' long-run outcomes than have been claimed.

—Jesse Rothstein is the Director of the Institute for Research on Labor and Employment and Professor in the Goldman School of Public Policy and the Department of Economics at the University of California, Berkeley. Contact: rothstein@berkeley.edu

Endnotes

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

[2] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104(9):2633-2679.

[3] Rothstein, Jesse. 2016. "Revisiting the Impact of Teachers," *Working Paper*. http://eml.berkeley.edu/~jrothst/CFR/rothstein_cfr.pdf.

[4] Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014b, *op. cit.*

[5] See p. 43 in Hanushek, Eric A. 2011. "Valuing Teachers: How Much is a Good Teacher Worth?" Education Next, 11(3):40-45.

[6] Other concerns center around the limited coverage of student tests, with the implication that VAM-based evaluations will lead teachers to focus on test-taking skills and on the content of the tests to the exclusion of other important domains and topics, and around the high volatility of teachers' VAM scores.

[7] Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1):175-214.

[8] See for example Kane, Thomas .J., and Douglas O. Staiger.2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *National Bureau of Economic Research Working Paper* #14607, and Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment", Report for the Bill & Melinda Gates Foundation, Seattle, WA.

[9] Chetty, Raj, John N. Friedman , and Jonah E. Rockoff. 2014a, *op. cit.*

[10] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b, *op. cit.*

[11] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2016. "Measuring the Impact of Teachers: Response to Rothstein (2014)", *Working Paper*.

[12] The non-randomness of measured teacher switching is in part attributable to the way that CFR measure value-added: They exclude teachers who are observed in the sample for only one or two years, and I show that this creates an association between VAM scores and student preparedness that leads CFR to overstate the accuracy of VAM-based predictions. Alternative strategies for measuring value-added for the excluded teachers reduce the association between teacher switching and student preparedness, while also leading to quite different overall conclusions.

[13] Chetty, Raj, John N. Friedman , and Jonah E. Rockoff. 2014b, *op. cit.*

The Institute for Research on Labor and Employment (IRLE) is an organized research unit at the University of California, Berkeley. Since its creation in 1945, IRLE has brought together faculty from across the university in support of multidisciplinary research about labor and employment. IRLE sponsors several community service programs and research centers, including the Center for Labor Research and Education (Labor Center), the Center for the Study of Child Care Employment, the Center on Wage and Employment Dynamics, the Don Vial Center on the Green Economy, California Public Employee Relations, and the Food Labor Research Center.

The policy brief series is edited by Claire Montialoux, Research Economist at IRLE. Contact: *claire.montialoux@berkeley.edu*

The complete IRLE policy brief series is available at **http://irle.berkeley.edu/policy-briefs/.**

Previous briefs include:

1. **The New California Earned Income Tax Credit**, by Claire Montialoux and Jesse Rothstein, Dec 2015.
2. **Can School Finance Reforms Improve Student Achievement?**, by Julien Lafortune, Jesse Rothstein, and Diane Schanzenbach, March 2016.
3. **The Supplemental Nutrition Assistance Program: A Central Component of the Social Safety Net**, by Hilary Hoynes, April 2016.
4. **Where Did All the Migrant Farm Workers Go?**, by Maoyong Fan and Jeffrey M. Perloff, July 2016.
5. **Revisiting the impact of Head Start**, by Claire Montialoux, September 2016.
6. **Is the Great Recession Really Over?**, by Danny Yagan, September 2016.
7. **Are Minimum Wage Increases Absorbed by Small Price Increases?**, by Sylvia Allegretto and Michael Reich, November 2016.
8. **Fiscal Policy and Employment: Lessons from the Social Security Earnings Test**, by Claire Montialoux, December 2016.