Michigan Reading Journal

Volume 48 | Issue 1 Article 11

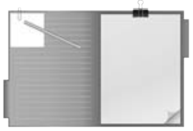
2016

Student Test Scores and Teacher Evaluation: What Do We Know and What Do We Need to Learn

Joshua M. Cowen

Student Test Scores and Teacher Evaluation: What Do We Know and What Do We Need to Learn?

by Joshua M. Cowen



Few recent issues of educational policy have generated so intense a debate as teacher evaluation. Of particular concern is the central place that standardized testing has taken in new laws changing the way schools and districts assess teacher effectiveness. In Michigan, for example, state lawmakers passed a comprehensive plan in 2011 that requires supervisors to include measures of student performance, where available, in evaluations of teacher performance. Supporters of these types of changes argue that student outcomes provide schools with "objective evidence" for success in the classroom (National Council on Teacher Quality, 2014). As Lily Eskelsen-Garcia, president of the National Education Association has put it, "using test scores is basically saying to educators, 'Hit your number or you get punished.' " Drawing on her own experience serving different students over time, Garcia explained, "Test scores alone wouldn't have told you what happened. They wouldn't have given you an analysis of why" (Bryant, 2014).

Critics like Garcia see testing and teacher evaluation as part of a larger "war on teachers." A more optimistic interpretation of these reforms is that policymakers are beginning to recognize what researchers, educators, and parents alike have long known: that effective teaching can make the difference between student success and failure, not just in school but beyond. Indeed, there is general agreement that teachers matter more than any other single school-based determinant of student outcomes—perhaps second only to outside-of-school factors like family background itself (e.g., Aaronson, Barrow, & Sander, 2007; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, et al., 2005; Rockoff, 2004; Chetty, Friedman, & Rockoff, 2013). We know that good teaching is associated not only with higher test scores for students, but also with higher later-in-life outcomes like college attendance and future salaries.

Joshua Cowen

The problem remains assessing effective teaching. Much of the same research has shown that few "observable" attributes of a teacher predict student success. In general, studies have found little evidence that teachers with Master's degrees in education are more effective than those with just an undergraduate degree; similarly, teacher certification tells us little about teacher effectiveness (Goldhaber & Brewer, 1997; Rivkin, et al. 2005; Kane, Rockoff & Staiger, 2008). Indeed, among the characteristics that school leaders can readily ascertain about a teacher, only experience appears correlated with student outcomes in multiple studies. More experienced teachers are more effective than less experienced teachers, but even here these differences do not persist forever. Most studies showing that experience "matters" tend to also find that after 5 to 8 years of teaching, experience no longer appears as important. This means that we cannot be sure that a teacher with 15 or 16 years is necessarily more effective than a teacher with only 9 or 10 years in the classroom (e.g. Clotfelter et al., 2010; Goldhaber & Hansen, 2013; Rivkin et al., 2005; Rockoff, 2004).

Implicit in this work is the idea that if those observable teacher characteristics like certification or years of education make little difference to student achievement, what "matters" are the particular unobservable attributes that each individual teacher brings into the classroom. This is one reason that direct observation—either by supervisors or peers—of teaching in the classroom has long been a part of professional development in most states. But the results of these classroom observations necessarily depend on who is doing the observing, and when. One assistant principal may have a different idea of what makes a good teacher than other teachers or administrators in the district, so educators observed by different supervisors may be evaluated on different criteria. Even teachers observed by the same person may appear more or less effective to that supervisor at different points in time—we all have good and bad days.

In the past decade, a number of experts have tried to address these problems by developing new ways to link individual teachers to differences in student outcomes. Perhaps the most controversial method involves a set of techniques known as "value-added models," or VAMs. Although based on complicated statistical methods about which many different experts still differ, the idea behind VAMs is largely intuitive. In essence, a VAM predicts what a student's test score should be based on a number of observable attributes about the student—race, gender, whether the student has special academic needs, whether he or she is eligible or participating in free/reduced lunch, or whether he or she is a non-native English speaker, to name a few major examples—as well as, usually, attributes about the school in which the student is learning. In each year, the difference between the student's actual test score and what is predicted by the VAM is attributed to the student's teacher. That teacher's "value added" score is essentially a summary of all of those differences between his/her students' actual and predicted outcomes. Because the most important student-level characteristic used in the VAM to predict a student's current test scores are that students' scores on the same or similar exams in the past, the ideal VAM may credibly be an estimate of a teacher's contributions to student learning even after having taken into account how prepared students were before entering his/her classroom.

The key is the word "may." A number of steps have to be taken for a VAM-based estimate of a teacher's effect to be credible. The first is whether the model uses the right information to predict a student's test score. What is the "right" information? In the VAM framework, any characteristic of a student that could be related to both that student's test score in a particular year and to whatever determined that student's assignment to a particular teacher must explicitly be included in the prediction. In practice, this is why the student's earlier tests are so important. If a principal is assigning more high-ability children to some teachers and more struggling students to others, failing to account for that pattern could lead to misattributing the former group of teachers as more effective than the latter. The good news is that for most students, prior test scores are indeed available, so this concern may be less of a problem than when researchers first raised it, although questions still remain (e.g. Rothstein, 2010; Koedel & Betts, 2011; Guarino, Reckase & Wooldridge, 2014).

Perhaps a more serious problem from the standpoint of tying VAM-based results to decisions about teachers' careers concerns the extent to which even valid estimates of a teacher's contribution to student learning remain imprecise. Think about the problem this way: if you have ever been ticketed by a traffic officer for speeding down the interstate, you may have been surprised, at first, to see flashing lights in your rearview mirror. This may be because you had a fairly good idea of your general speed—you know you weren't traveling at 35 miles per hour, for example, but you are also confident that you were well below 90. Whether your actual speed was 70 or 80 miles per hour, however—a much narrower range than 35 to 90—probably determined the officer's decision to pull you over. The exact recorded speed will also determine the financial cost of your ticket (if you are 15 m.p.h. over the limit you'll pay more than if you are under 10) as well as any punitive points added to your license. The problem is similar in teacher evaluation. Unfortunately, as helpful as VAMs can be, they are not as reliable as the sophisticated radars employed by the Michigan State Police (although it has always been my bad luck to drive by officers with faulty detectors). This means that, even if in the best case scenario, administrators can be confident that a VAM-based system adequately separates highly effective from highly ineffective instructors, where to draw the line between particular categories of teachers—say, the difference between a truly ineffective teacher and one who is "only" below average is much more difficult.

The traffic example is also informative because, as with cut-points used to determine what is a safe road speed and what is not, the categories of "effective" and "ineffective" used to make decisions about a teacher's tenure case or whether to retain the teacher at all are ultimately a matter of policy. And for individual teachers near to the "below average" and "ineffective" cutoff, the difference between being on one side and the other could have profound career consequences. One way to get around this problem is to consider repeated measures of a teacher's effectiveness before adding consequences that could result in dismissal. Another is to build multiple measures of performance into each year's final rating. For example, a teacher with a relatively low VAM score may still have higher scores assigned by his or her classroom observer. In nearly all states that have begun using test scores to evaluate teachers, classroom observations remain an integral part of teacher assessment (National Council on Teacher Quality, 2014).

The key feature of Michigan's version of these changes, which began in 2011, was the inclusion of student achievement as a "significant" determinant of educator performance ratings, and the eventual dismissal of teachers with multiple (three) "ineffective" ratings. Until 2013-2014, districts were allowed to establish their own definition of "significant," after which time at least 25% of teachers' overall scores were to be determined by student outcomes, with at least 50% of over all ratings determined by test scores in future years (Michigan Department of Education, 2014; State of Michigan, 2011). This plan originally called for the implementation of a statewide system of teacher evaluation, the details of which to be recommended by a team of experts across the state. That team's recommendations came back more than two years ago (Michigan Council for Educator Effectiveness, 2013), although a statewide system has yet to be in place.

There are a number of advantages to such a statewide system over one that is locally based. The first might be called operational. Most individual districts may be unable to meet the difficulty of gathering data over time or the challenges of developing and implementing the statistical techniques necessary to take advantage of the good aspects of VAMs while avoiding some of the problems. Another concerns jurisdiction. Since states set certification criteria and determine laws like teacher tenure, it makes sense that criteria for teacher evaluation should be uniform across a particular state—or at least uniform for teachers in the same grade and subject. The final advantage is the basic and related issue of fairness. Only in a state system can a teacher be sure that the criteria used to create his or her VAM in one district are largely the same as those creating VAM ratings for teachers elsewhere. The downside is that the needs of individual districts may vary, and a statewide system may address these local needs inadequately. Such a tradeoff was present in recent efforts by members of the Michigan legislature to forestall the planned statewide system (French, 2015).

Moreover, there are other issues that even a well-designed VAM-based teacher evaluation system—whether at the state or local level—will be unable to solve. The simple matter of verifying rosters of students for a given teacher in a given year is difficult, especially in areas with high rates of student mobility, but such verification is essential to ensuring teachers are evaluated based on the results of children they actually teach. Recent indications from Tennessee's system have underscored the difficulties in properly tracking which children are assigned to which teachers over time (Springer & Ballou, 2015). There is also the related problem of separating individual teacher effects from more collaborative efforts across a particular grade or school. Although there are mechanical ways to address such a difficulty in a particular VAM model, statistics give little guide to the proper way to weigh individual from group contributions to student success. In addition, any system based on student test scores—whether VAM or otherwise—is only as good as the tests themselves. This limitation is most clearly present in the extreme case of teachers in subjects that are untested (in which case other methods of assessing effects on student outcomes are required), but is from a more philosophical standpoint also evident in the question of whether a standardized test truly reflects student knowledge or aptitude. Although few serious proponents of VAM-based evaluations would claim, as one reform critic has charged they do, that these models "factor out things such as a student's intelligence, whether the student is hungry, sick or is subject to violence at home" (Strauss, 2015), implicit in policies that employ these methods is indeed the assumption that a teacher's chief responsibility is cultivating student skills that are readily measureable by tests. Faulting VAMs for failing to account for other, myriad ways that teachers make a difference in the lives of their students misplaces the blame on the tool instead of its user.

Where does all of this leave policymakers, practitioners, and parents? Experts will say, as always, that more research on VAMs and other forms of teacher evaluation is needed. And this is undoubtedly true, although many of the major strengths and limitations of VAM-based approaches are already well-understood (Corcoran & Goldhaber, 2013). Less recognized, and much less accepted, are the implicit tradeoffs required in any system of teacher evaluation—including the uncomfortable reality that what is fair for individual teachers may not always be what is best for the students they serve (Goldhaber, 2015). Resolving these tensions goes well beyond statistics or philosophies of student learning. In a democracy, how we evaluate our teachers, and on what basis, are ultimately matters of civic responsibility and engagement.

References

Aaronson, D.,Barrow, L., & Sander, W. (2007) Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics 25*(1), 95-135.

Bryant, J.. (2014) Stupid, absurd, non-defensible: new president Lily Eskelsen-Garcia on the problem with Arne Duncan, standardized tests and the war on teachers. Retrieved from http://www.salon.com/2014/07/30/stupid_absurd_non_defensible_new_nea_president_lily_eskelsen_garcia_on_the_problem_with_arne_duncan_standardized_tests_and_the_war_on_teachers/

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy 8*(3), 418-434.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood (Working Paper No. 19424). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19424

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

French, R.. (2015). "How a single powerful senator killed serious reform of teacher evaluation." *Bridge Magazine* 6/4/15. Retrieved from http://bridgemi.com/2015/06/how-a-single-powerful-senator-killed-serious-reform-of-teacher-evaluation/

Guarino, C. M., Reckase, M.D., &Wooldridge, J.M. (2014) Can value-added measures of teacher performance be trusted? *Education Finance and Policy 10*(1), 117-156

Goldhaber, D.D., &Brewer, D.J. (1997) Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources 32*(3), 505-523.

Goldhaber, D., &Hansen, M. (2013) Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica 80*(319), 589-612.

Goldhaber, D. (2015) Exploring the Potential of Value-Added Performance Measures to Affect the Quality of the Teacher Workforce. *Educational Researcher 44*(2), 87–95.

Kane, T. J., = Rockoff, J.E., &Staiger, D.O.. (2008) What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review 27*(6).

Koedel, C., & Betts, J. R.. (2011) Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Polioсy 6*(1), 18-42.

Michigan Council for Educator Effectiveness (2013). Building an Improvement-Focused System of Educator Evaluation in Michigan: Final Recommendations. Retrieved fromfile:///Users/jcowen/Downloads/midnightreport_july24_2013.pdf

Michigan Department of Education (2014). *Educator Evaluation and Effectiveness in Michigan 2013-2014.* Lansing, Michigan

National Council on Teacher Quality (2014) State Policy Yearbook; Policy

Issue: Teacher Evaluation. Retrieved fromhttp://www.nctq.org/statePolicy/2014/policyIssueFindings.do?policyIssueId=6&masterGoalId=11&stateId=23&yearId=7

Nye, B., Konstantopoulos, S., &Hedges, L.V. (2004) "How large are teacher effects?." *Educational Evaluation and Policy Analysis 26*(3), 237-25.

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

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

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

Springer, M. G., &Ballou, D. (2015) Using Student Test Scores to Measure Teacher Performance: Some Problems in the Design and Implementation of Evaluation Systems. *Educational Researcher 44*(2), 77–86.

State of Michigan (2011) 96th *Regular Session of 2011 Public Acts 10* Retrieved from https://www.legislature.mi.gov/documents/2011-2012/publicact/htm/2011-PA-0102.htm

Author Biography

Joshua Cowen, Ph.D. is an Associate Professor of Education Policy at Michigan State University. His research focuses broadly on school choice and teacher quality. For questions, comments, or additional information please contact him at jcowen@msu.edu.