BROWN CENTER on
Education Policy
at BROOKINGS

November 17, 2010

Evaluating Teachers:
The Important Role of Value-Added

Yellow Dog Productions

The Brookings Brown Center Task Group on Teacher Quality

Steven Glazerman, Mathematica Policy Research
Susanna Loeb, Stanford University
Dan Goldhaber, University of Washington
Douglas Staiger, Dartmouth University
Stephen Raudenbush, University of Chicago
Grover Whitehurst, The Brookings Institution

EXECUTIVE SUMMARY

The evaluation of teachers based on the contribution they make to the learning of their students, *value-added*, is an increasingly popular but controversial education reform policy. We highlight and try to clarify four areas of confusion about value-added. The first is between value-added information and the uses to which it can be put. One can, for example, be in favor of an evaluation system that includes value-added information without endorsing the release to the public of value-added data on individual teachers. The second is between the consequences for teachers vs. those for students of classifying and misclassifying teachers as effective or ineffective — the interests of students are not always perfectly congruent with those of teachers. The third is between the reliability of value-added measures of teacher performance and the standards for evaluations in other fields — value-added scores for individual teachers turn out to be about as reliable as performance assessments used elsewhere for high stakes decisions. The fourth is between the reliability of teacher evaluation systems that include value-added vs. those that do not — ignoring value-added typically lowers the reliability of personnel decisions about teachers. We conclude that value-added data has an important role to play in teacher evaluation systems, but that there is much to be learned about how best to use value-added information in human resource decisions.

There is an obvious need for teacher evaluation systems that include a spread of verifiable and comparable teacher evaluations that distinguish teacher effectiveness.

Teacher evaluation at a crossroads

The vast majority of school districts presently employ teacher evaluation systems that result in all teachers receiving the same (top) rating. This is perhaps best exemplified by a recent report by the New Teacher Project focusing on thousands of teachers and administrators spanning twelve districts in four states.[1] The report revealed that even though all the districts employed some formal evaluation process for teachers, all failed to differentiate meaningfully among levels of teaching effectiveness. In districts that used binary ratings more than 99 percent of teachers were rated satisfactory. In districts using a broader range of ratings, 94 percent received one of the top two ratings and less than 1 percent received an unsatisfactory rating. As Secretary of Education Arne Duncan put it, "Today in our country, 99 percent of our teachers are above average."[2]

There is an obvious need for teacher evaluation systems that include a spread of verifiable and comparable teacher evaluations that distinguish teacher effectiveness. We know from a large body of empirical research that teachers

[1] Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.
[2] Gabriel, T. (2010, September 2). A celebratory road trip by education secretary, *New York Times*, p. A24.

differ dramatically from one another in effectiveness. Evaluation systems could recognize these differences but they generally don't. As a consequence, the many low stakes and high stakes decisions that are made in the teacher labor market occur without the benefit of formalized recognition of how effective (or ineffective) teachers are in the classroom. Is there any doubt that teacher policy decisions would be better informed by teacher evaluation systems that meaningfully differentiate among teachers?

There is tremendous support at both the federal and state levels for the development and use of teacher evaluation systems that are more discerning.[3] And the two national teachers unions, the AFT and the NEA, support teacher evaluation systems that recognize and reward excellence and improve professional development. This is consistent with their long-term support of the National Board for Professional Teaching Standards, which is designed to identify excellent teachers and provide them a salary bonus.

The latest generation of teacher evaluation systems seeks to incorporate information on the value-added by individual teachers to the achievement of their students. The teacher's contribution can be estimated in a variety of ways, but typically entails some variant of subtracting the achievement test score of a teacher's students at the beginning of the year from their score at the end of the year, and making statistical adjustments to account for differences in student learning that might result from student background or school-wide factors outside the teacher's control. These adjusted *gains* in student achievement are compared across teachers. Value-added scores can be expressed in a number of ways. One that is easy to grasp is a percentile score that indicates where a given teacher stands relative to other teachers. Thus a teacher who scored at the 75th percentile on value-added for mathematics achievement would have produced greater gains for her students than the gains produced by 75 percent of the other teachers being evaluated.

Critics of value-added methods have raised concerns about the statistical validity, reliability, and corruptibility of value-added measures. We believe the correct response to these concerns is to improve value-added measures continually and to use them wisely, not to discard or ignore the data. With that goal in mind, we address four sources of concern about value-added evaluation of teachers

Value-added information vs. what you do with it

There is considerable debate about how teacher evaluations should be used to improve schools, and uncertainty about how to implement proposed reforms. For example, even those who favor linking pay to performance face numerous design

[3] For instance, the Obama administration made state support of rigorous teacher evaluation systems a pre-condition for competition in Race to the Top, and has laid out a blueprint for the reauthorization of the Elementary and Secondary Education Act in which teacher effectiveness defined by evaluation of on-the-job performance is an important facet.

decisions with uncertain consequences. How a pay for performance system is designed—salary incentives based on team performance vs. individual performance, having incentives managed from the state or district level vs. the building level, or having incentives structured as more rapid advancement through a system of ranks vs. annual bonuses—can result in very good or very ineffective policy.[4]

Similar uncertainty surrounds other possible uses of value-added information. For example, tying tenure to value-added evaluation scores will have immediate effects on school performance that have been well modeled, but these models cannot predict indirect effects such as those that might result from changes in the profiles of people interested in entering the teaching profession. Such effects on the general equilibrium of the teacher labor market are largely the subject of hypothesis and speculation. Research on these and related topics is burgeoning,[5] but right now much more is unknown than known.

Much of the controversy surrounding teacher performance measures that incorporate value-added information is based on fears about how the measures will be used.

However, uncertainties surrounding how best to design human resource policies that take advantage of meaningful teacher evaluation do not bear directly on the question of whether value-added information should be included as a component of teacher evaluation. There is considerable confusion between issues surrounding the inclusion of value-added scores in teacher evaluation systems and questions about how such information is used for human resource decisions. This is probably because the uses of teacher evaluation that have gained the most public attention or notoriety have been based *exclusively* on value-added. For example, in August 2010, the *Los Angeles Times* used several years of math and English test data to identify publicly the best and the worst third- to fifth-grade teachers in the Los Angeles Unified School District. The ensuing controversy focused as much on value-added evaluation as the newspaper's actions. But the question of whether these kinds of statistics should be published is separable from the question of whether such data should have a role in personnel decisions. It is routine for working professionals to receive consequential evaluations of their job performance, but that information is *not* typically broadcast to the public.

A place for value-added

Much of the controversy surrounding teacher performance measures that incorporate value-added information is based on fears about how the measures will be used. After all, once administrators have ready access to a quantitative performance measure, they can use it for sensitive human resources decisions including teacher pay, promotion, or layoffs. They may or may not do this wisely

[4] Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

[5] Goldhaber, D. & Hannaway, J. (Eds.) (2009). *Creating a new teaching profession*. Washington, DC: The Urban Institute.

or well, and it is reasonable for those who will be affected to express concerns.

We believe that whenever human resource actions are based on *evaluations* of teachers they will benefit from incorporating all the best available information, which includes value-added measures. Not only do teachers typically receive scant feedback on their past performance in raising test scores, the information they usually receive on the average test scores or proficiency of their students can be misleading or demoralizing. High test scores or a high proficiency rate may be more informative of who their students are than how they were taught. Low test scores might mask the incredible progress the teachers made. Teachers and their mentors and principals stand to gain vast new insight if they could see the teachers' performance placed in context of other teachers with students just like their own, drawn from a much larger population than a single school. This is the promise of value-added analysis. It is not a perfect system of measurement, but it can complement observational measures, parent feedback, and personal reflections on teaching far better than any available alternative. It can be used to help guide resources to where they are needed most, to identify teachers' strengths and weaknesses, and to put a spotlight on the critical role of teachers in learning.

Full-throated debate about policies such as merit pay and "last in-first out" should continue, but we should not let controversy over the uses of teacher evaluation information stand in the way of developing and improving measures of teacher performance.

Some classification errors are worse than others

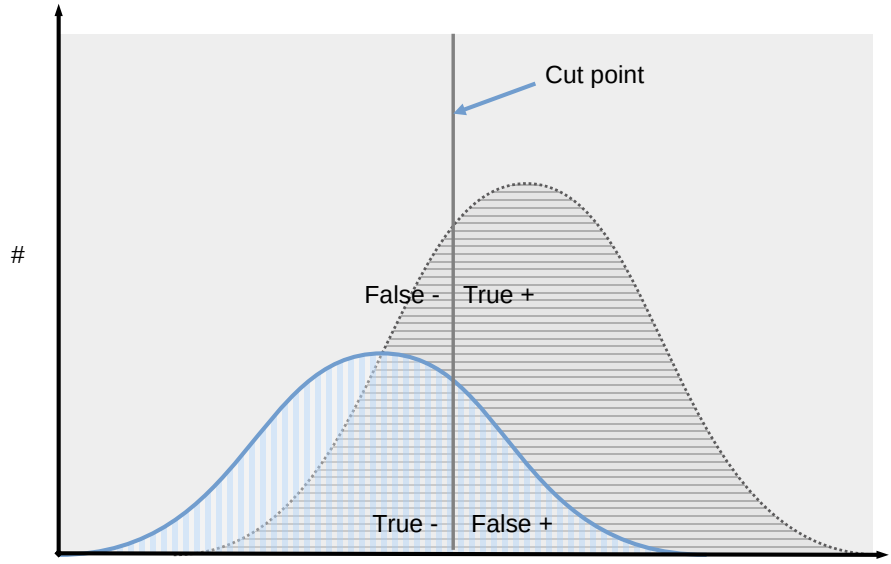
Recent reports by nationally visible education researchers and thinkers have urged restraint in the use of teacher evaluations based on student test scores for high stakes decisions. The common thread in these reports is the concern that value-added scores reported at the level of individual teachers frequently misclassify teachers in ways that are unfair to teachers, e.g., identifying a teacher as ineffective who is in fact average.[6]

There are three problems with these reports. First, they often set up an impossible test that is not the objective of any specific teacher evaluation system, such as using a single year of test score growth to produce a rank ordered list of

[6] For example, a policy brief from the Education Policy Institute on the problems with the use of student test scores to evaluate teachers, reports that value-added estimates "have proven to be unstable across statistical models, years, and classes that teachers teach." The authors, buttress their recommendations not to use such scores with descriptions of research showing that "among teachers who were ranked in the top 20 percent of effectiveness in the first year, fewer than a third were in that top group the next year," and that "effectiveness ratings in one year could only predict from 4 percent to 16 percent of the variation in such ratings in the following year." And, a report from the National Academies of Science presents a range of views on the use of value-added but nevertheless concludes that "persistent concerns about precision and bias militate against employing value-added indicators as the principal basis for high-stakes decisions." Likewise, reports from Rand, the Educational Testing Service, and IES remind us to be cautious about the degree of precision in estimates of teacher effectiveness derived from value-added measures.

teachers for a high stakes decision such as tenure. Any practical application of value-added measures should make use of confidence intervals in order to avoid false precision, and should include multiple years of value-added data in combination with other sources of information to increase reliability and validity. Second, they often ignore the fact that all decision-making systems have classification error. The goal is to minimize the most costly classification mistakes, not eliminate all of them. Third, they focus too much on one type of classification error, the type that negatively affects the interests of individual teachers.

Imagine the simplest classification system that could be fit on a continuous distribution of teachers' value-added scores: A point on the distribution is selected as a cut point. Any teacher receiving a value-added score at or above that cut point is categorized as effective whereas any teacher with a score below that point is categorized as ineffective. Imagine further that value-added is measured with error, i.e., a teacher's score does not capture perfectly the teacher's true contribution to student learning. This error in measurement means that depending on where the cut point is placed, some truly effective teachers will be rated ineffective (they are false negatives) and some ineffective teachers will be rated effective (they are false positives). The other two classification outcomes are truly effective teachers so categorized (true positives), and truly ineffective teachers so categorized (true negatives).



To illustrate, the figure above represents the obtained evaluation scores of two categories of teachers: those who are truly effective (colored grey) and those who are truly ineffective (colored blue). The scores of the two groups of teachers are distributed normally around the mean for their group, with the spread of scores representing both true differences in teacher effectiveness and error in the measure

used for evaluation. The cut point in the figure represents the point on the scale of teacher evaluation scores at which a manager chooses to treat the teachers differently in terms of a personnel action. Using tenure as an example, everyone who received an evaluation score at or above the cut point would receive tenure, whereas everyone scoring below the cut point would be dismissed or continue in a probationary status. In this instance, the majority of truly effective teachers received scores at or above the cut point – they are true positives – and a majority of truly ineffective teachers received scores below the cut point – they are true negatives. But there are also classification errors, i.e., truly effective teachers categorized as ineffective (false negatives) and truly ineffective teachers classified as effective (false positives).

The false positive rate and the false negative rate are inversely related and determined by where the cut point is placed on the distribution of scores. Thus, if the manager moved the cut point for granting tenure to the right in this figure, the false positive rate would go down whereas the false negative rate would go up. Likewise the true positive rate would go up and the true negative rate would go down.

An evaluation system that results in tenure and advancement for almost every teacher and thus has a very low rate of false negatives generates a high rate of false positives, i.e., teachers identified as effective who are not.

Much of the concern and cautions about the use of value-added have focused on the frequency of occurrence of false negatives, i.e., effective teachers who are identified as ineffective. But framing the problem in terms of false negatives places the focus almost entirely on the interests of the individual who is being evaluated rather than the students who are being served. It is easy to identify with the *good* teacher who wants to avoid dismissal for being incorrectly labeled a *bad* teacher. From that individual's perspective, no rate of misclassification is acceptable. However, an evaluation system that results in tenure and advancement for almost every teacher and thus has a very low rate of false negatives generates a high rate of false positives, i.e., teachers identified as effective who are not. These teachers drag down the performance of schools and do not serve students as well as more effective teachers.

In the simplest of scenarios involving tenure of novice teachers, it is in the best interest of students to raise the cut point thereby increasing the proportion of truly effective teachers staffing classrooms whereas it is in the best interest of novice teachers to lower the cut point thereby making it more likely that they will be granted tenure. Our message is that the interests of students and the interests of teachers in classification errors are not always congruent, and that a system that generates a fairly high rate of false negatives could still produce better outcomes for students by raising the overall quality of the teacher workforce.[7] A focus on the

[7] Of course, there are many tradeoffs that belie the simple calculus in our example. For instance, if an appreciable share of junior teachers were removed from the workforce in a particular district the pool of applicants might be too small to replace the dismissed teachers. From a district or student's perspective it would be better to have lower quality teachers in the classroom than no teachers at all. Likewise, the calculus is not straightforward from a teacher's perspective. For example an evaluation system that identifies nearly everyone as a winner and thereby avoids false negatives may lessen the opportunities for advancement of stronger teachers and reduce the public's support for the teaching profession.

effects on teachers of misclassification should be balanced by a concern with the effects on students.

A performance measure needs to be good, not perfect

Discussions of teacher evaluation at the policy and technical levels often proceed in isolation from experience and evidence from other related fields. But we know a lot about performance evaluation in other labor markets, knowledge that should inform debates about value-added and teacher evaluation in general.

The correlation in test-based measures of teaching effectiveness between one school year and the next lies between .20 and .60 across multiple studies, with most estimates lying between .30 and .40.[8] A measure that has a correlation of .35 from one year to the next produces seemingly troubling statistics in line with our conceptual discussion of classification errors. For instance, only about a third of teachers ranked in the top quartile of value-added based on one academic year's performance would appear in the top quartile again the next year. And ten percent of bottom quartile teachers one year would appear in the top quartile the next. Some of this instability is due to variation in teachers' true performance from year to year and some of it is simply due to error in the measure.

It is instructive to look at other sectors of the economy as a gauge for judging the stability of value-added measures. The use of imprecise measures to make high stakes decisions that place societal or institutional interests above those of individuals is wide spread and accepted in fields outside of teaching.

The correlation of the college admission test scores of college applicants with measures of college success is modest (r = .35 for SAT combined verbal + math and freshman GPA[9]). Nevertheless nearly all selective colleges use SAT or ACT scores as a heavily weighted component of their admission decisions even though that produces substantial false negative rates (students who could have succeeded but are denied entry). Why would colleges use such a flawed selection instrument? Because even though the prediction of success from SAT/ACT scores is modest it is among the strongest available predictors. An entering class formed in part by the decision to admit those with higher SAT/ACT scores in preference to those with lower scores will perform better than a class formed without the use of that information.

In health care, patient volume and patient mortality rates for surgeons and hospitals are publicly reported on an annual basis by private organizations and federal agencies and have been formally approved as quality measures by national

[8] Goldhaber, D. & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance*. CEDR Working Paper 2010-3. Seattle, WA: University of Washington.

[9] Camera, W.J. & Echternacht, G. (July 2000). *The SAT I and high school grades: Utility in predicting success in college*. New York, NY: The College Board.

organizations.[10] Yet patient volume is only modestly correlated with patient outcomes, and the year-to-year correlations in patient mortality rates are well below 0.5 for most medical and surgical conditions. Nevertheless, these measures are used by patients and health care purchasers to select providers because they are able to predict larger differences across medical providers in patient outcomes than other available measures.[11]

In a similar vein, the volume of home sales for realtors; returns on investment funds; productivity of field-service personnel for utility companies; output of sewing machine operators; and baseball batting averages predict future performance only modestly. A meta-analysis[12] of 22 studies of objective performance measures found that the year-to-year correlations in high complexity jobs ranged from 0.33 to 0.40, consistent with value-added correlations for teachers.

Rather than asking value-added to measure up to an arbitrary standard of perfection, it would be productive to ask how it performs compared to classification based on other forms of available information of teachers.

Despite these modest predictive relationships, real estate firms rationally try to recruit last year's volume leader from a competing firm; investors understandably prefer investment firms with above average returns in a previous year; and baseball batting averages in a given year have large effects on player contracts. The between-season correlation in batting averages for professional baseball players is .36.[13] Ask any manager of a baseball team whether he considers a player's batting average from the previous year in decisions about the present year.

We should not set unrealistic expectations for the reliability or stability of value-added. Value-added evaluations are as reliable as those used for high stakes decisions in many other fields.

Ignoring value-added data doesn't help

We know a good deal about how other means of classification of teachers perform versus value-added. Rather than asking value-added to measure up to an arbitrary standard of perfection, it would be productive to ask how it performs compared to classification based on other forms of available information of teachers.

The "compared to what" question has been addressed by a good deal of research on the other teacher credentials and characteristics that are presently used to determine employment eligibility and compensation. Here the research is quite

[10] See http://www.leapfroggroup.org/, http://www.hospitalcompare.hhs.gov/, and http://www.qualityforum.org/Measures_List.aspx.

[11] For example, Dimick, J.B., Staiger, D.O., Basur, O., & Birkmeyer, J.D. (2009). Composite measures for predicting surgical mortality in the hospital. *Health Affairs, 28(4),* 1189-1198.

[12] Sturman, M.C., Cheramie, R.A., & and Cashen, L.H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, *90*, 269-283.

[13] Schall, T. & Smith, G. (2000). Do baseball players regress to the mean? *The American Statistician*, *54*, 231-235.

clear: if student test achievement is the outcome,[14] value-added is superior to other existing methods of classifying teachers. Classification that relies on other measurable characteristics of teachers (e.g., scores on licensing tests, routes into teaching, nature of certification, National Board certification, teaching experience, quality of undergraduate institution, relevance of undergraduate coursework, extent and nature of professional development), considered singly or in aggregate, is not in the same league in terms of predicting future performance as evaluation based on value-added.

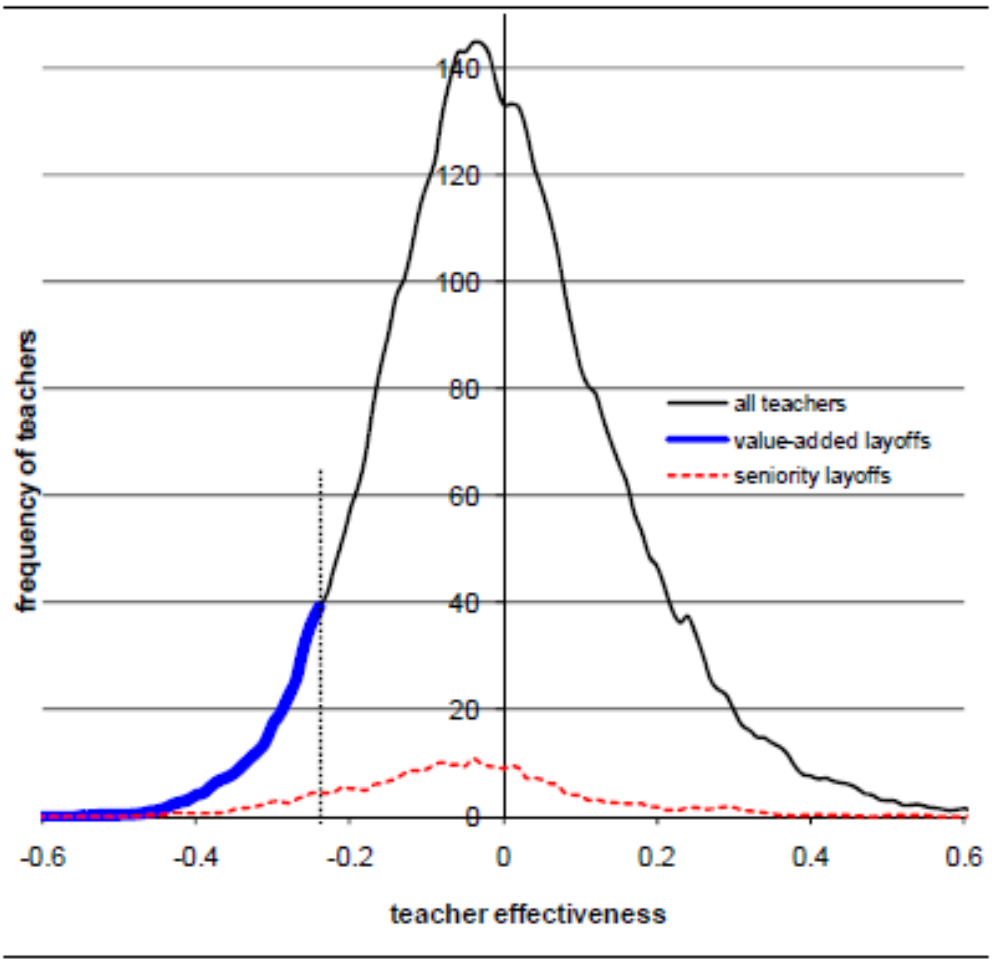
Consider a particular example that has arisen as a consequence of the deep recession: the need of districts to lay off teachers as a result of budget shortfalls. Managers in most industries would attempt to target layoffs so as to cause as little damage as possible to productivity — less productive workers would be dismissed or furloughed before more productive workers.

Suppose school district leaders were similarly motivated and had flexibility in deciding how to proceed. Imagine three possible approaches for deciding who should be dismissed. The first approach would employ the existing teacher evaluation system based on principal ratings, which identifies a few teachers as unsatisfactory but categorized the vast majority of teachers as satisfactory. The second approach would employ teacher experience, which has been found in a number of studies to have a statistically significant positive association with student achievement. The third approach would use teacher value-added scores to identify the lowest performing teachers.

Researchers have compared these three approaches using data from fourth and fifth grade public school teachers in New York City and simulating the elimination of enough teachers to reduce the budget by 5 percent.[15] A graph from that study, reproduced below, illustrates the results for student achievement if the positions of teachers with the lowest value-added scores were eliminated vs. the positions of teachers with the least experience. The horizontal axis is teacher effectiveness as indexed by student gains whereas the vertical axis is the number of teachers. Teacher effectiveness scores are those regularly calculated by the NYC public schools and could encompass teacher performance going back as far as four years.

[14] Although student scores on standardized achievement tests are obviously proxies for rather than the actual student outcomes that education is supposed to generate, it is important to remember that they are strong predictors of long term outcomes. For example, a large scale national study by the ACT found that eighth-grade achievement test scores were the best predictor of students' level of college and career readiness at high school graduation —even more so than students' family background, high school coursework, or high school grade point average.

[15] Boyd, D.J., Lankford, H., Loeb, S., & Wyckoff, J.H. (July, 2010). *Teacher layoffs: An empirical illustration of seniority vs. measures of effectiveness*. Brief 12. National Center for Evaluation of Longitudinal Data in Education Research. Washington, DC: The Urban Institute.



Note that if teachers were laid off based on seniority they would be distributed across the full range of performance in terms of effectiveness in raising student test scores whereas teachers laid off based on low value-added scores would be at the bottom of the distribution. In other words, many more effective teachers would be retained were layoffs based on value-added than were they based on seniority. Principal ratings, not shown in the graph, perform better than teacher seniority in identifying teachers with low effectiveness in raising student achievement, but not nearly as well as value-added scores.

The question, then, is not whether evaluations of teacher effectiveness based on value-added are perfect or close to it: they are not. The question, instead, is whether and how the information from value-added compares with other sources of information available to schools when difficult and important personnel decisions must be made. For example, keeping ineffective teachers on the job while dismissing far better teachers is something most school leaders, parents, and the general public would want to avoid. Value-added is a better tool for that

purpose than other measures such as teacher experience, certification status, seniority, and principal ratings, even though it is imperfect.[16]

Conclusion: Value-added has an important role to play

We have a lot to learn about how to improve the reliability of value-added and other sources of information on teacher effectiveness, as well as how to build useful personnel policies around such information. However, too much of the debate about value-added assessment of teacher effectiveness has proceeded without consideration of the alternatives and by conflating objectionable personnel policies with value-added information itself. When teacher evaluation that incorporates value-added data is compared against an abstract ideal, it can easily be found wanting in that it provides only a fuzzy signal. But when it is compared to performance assessment in other fields or to evaluations of teachers based on other sources of information, it looks respectable and appears to provide the best signal we've got.

Teachers differ dramatically in their performance, with large consequences for students. Staffing policies that ignore this lose one of the strongest levers for lifting the performance of schools and students. That is why there is great interest in establishing teacher evaluation systems that meaningfully differentiate performance.

Teaching is a complex task and value-added captures only a portion of the impact of differences in teacher effectiveness. Thus high stakes decisions based on value-added measures of teacher performance will be imperfect. We do not advocate using value-added measures alone when making decisions about hiring, firing, tenure, compensation, placement, or developing teachers, but surely value-added information ought to be in the mix given the empirical evidence that it predicts more about what students will learn from the teachers to which they are assigned than any other source of information.

[16] Research related to this conclusion includes:

Goldhaber, D. D. & Hansen, M. (2009). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Working Paper 2009-2. Seattle, WA: Center on Reinventing Public Education.

Jacob, B. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*. 26(1), 101-36.

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), 615-31.

About the Brown Center on Education Policy

Established in 1992, the Brown Center on Education Policy conducts research and provides policy recommendations on topics in American education. The Brown Center is part of The Brookings Institution, a private nonprofit organization devoted to independent research and innovative policy solutions. For more than 90 years, Brookings has analyzed current and emerging issues and produced new ideas that matter - for the nation and the world.

Brown Center on Education Policy

The Brookings Institution
1775 Massachusetts Ave. NW
Washington DC, 20036
202.797.6090
202.797.6144 (f)
http://www.brookings.edu/brown.aspx

We would like to thank the Walton Family Foundation, the Foundation for Educational Choice, and an anonymous foundation for funding the Rethinking the Federal Role in Education project.

Governance Studies
The Brookings Institution
1775 Massachusetts Ave., NW
Washington, DC 20036
Tel: 202.797.6090
Fax: 202.797.6144
www.brookings.edu/governance.aspx

Editor
Christine Jacobs

Production & Layout
John S Seo

Email your comments to gscomments@brookings.edu