Commentary

Value-Added: It's Not Perfect, But It Makes Sense

By Steven Glazerman, Dan Goldhaber, Susanna Loeb, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst

The vast majority of school districts presently employ teacher-evaluation systems that result in nearly all teachers’ receiving the same (top) rating. For instance, a recent study of 12 districts in four states by the New Teacher Project revealed that more than 99 percent of teachers in districts using binary ratings were rated satisfactory, while 94 percent received one of the top two ratings in districts using a broader range of ratings. As U.S. Secretary of Education Arne Duncan put it during his bus tour this fall, “Today in our country, 99 percent of our teachers are above average.”

The reality is far different from what the evaluation systems suggest. We know from a large body of empirical research that teachers differ dramatically from one another in effectiveness. That today’s evaluation systems fail to recognize these differences means that the many important human-resource decisions are not as efficient or fair as they could be if they incorporated data that meaningfully differentiated among teachers.

Newer teacher-evaluation systems seek to incorporate information about individual teachers based on value-added measures of a teacher’s contribution toward student achievement. The teacher’s contribution can be estimated in a variety of ways, but typically entails some variant of subtracting the achievement-test scores of a teacher’s students at the beginning of the year from their scores at the end of the year, and making statistical adjustments to account for differences in student learning that might result from student background or schoolwide factors outside the teacher’s control. These adjusted gains in student achievement are compared across teachers.

Researchers have pointed out that value-added estimates for individual teachers fluctuate from year to year and can be influenced by factors over which the teacher has no control. The technical issues that have been raised about value-added measures would arise in one form or another with respect to any evaluation of complex human behavior. We believe the correct response to these concerns is to improve value-added measures continually and use them wisely. We should not discard or ignore the information they contain. With that goal in mind, we address four frequently cited concerns about the value-added evaluation of teachers.
• The Use of Value-Added Information

Much of the controversy surrounding teacher-performance measures that incorporate value-added is based on fears about how the information will be used. After all, once administrators have ready access to a quantitative performance measure, they can use it for such sensitive human-resource decisions as teacher pay, promotion, and layoffs. Administrators may or may not do this wisely or well, and it is reasonable for those who will be affected to express concerns.

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 on teachers.

We believe that whenever human-resource actions are based on evaluations of teachers, they will benefit from incorporating all the available information that improves prediction of student outcomes, which includes value-added measures. Full-throated debate and research on policies such as merit pay and “last in, first out” layoffs 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.

• Trading Classification Errors to Benefit Students

The common thread in technical critiques of value-added evaluation is that teachers subjected to it will often be misclassified, e.g., a teacher who is identified as “ineffective” is, in fact, “average.” Given the typical reliability of value-added measures, there is no doubt that such misclassifications will occur with some frequency. However, we must recognize that all decision making systems generate classification errors, including those used today. Moreover, different types of errors have different consequences.

In the case of teacher value-added, the focus has been almost entirely on so-called false-negative errors, i.e., teachers who are falsely classified as ineffective because the measures are not perfectly reliable. But framing the problem in terms of false negatives places the focus almost entirely on the interests of the teacher who is being evaluated rather than the students who are being served.

In the simplest of scenarios involving tenure of novice teachers, it is in the best interest of students to have a high bar set for effectiveness, thereby increasing the proportion of truly effective teachers staffing classrooms (i.e., fewer false positives); by contrast, it is in the best interest of novice teachers to have a low bar set for effectiveness, thereby making it more likely that they will be granted tenure (i.e., fewer false negatives). The administrator must trade off one type of classification error for the other when deciding how high to set the cut score for effectiveness based on teacher-evaluation scores.

We believe that the concern with the effects of misclassification on teachers should be balanced by a concern with the effects on students.
• The Setting of Realistic Value-Added Benchmarks

The correlation of value-added 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. A measure that has a correlation of .35 from one year to the next will result in a significant number of classification errors, consistent with our previous point. But is the amount of error in classification too high to be tolerated?

It is instructive to look at other sectors of the economy as a gauge for judging whether value-added measures are sufficiently stable to be used for high-stakes decisions. 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 organizations. Yet patient volume is only modestly correlated with patient outcomes, and the year-to-year correlations in patient-mortality rates are well below .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 are.

In a similar vein, the volume of home sales for real estate agents, returns on investment funds, college-entrance examinations, 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 of 22 studies of objective performance measures found that the year-to-year correlations in high-complexity jobs ranged from .33 to .40, consistent with value-added correlations for 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; colleges select students with higher entrance-exam scores; 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. Ask any manager of a baseball team whether a player’s batting average from the previous year is relevant in making decisions about the present year.

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

• The Reliability of Value-Added as a Measurement of Effectiveness

We know a good deal about how other means of classification of teachers perform vs. 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 on teachers.

Here the research is quite clear: If student test achievement is the desired outcome, 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, the path to certification, National Board for Professional Teaching Standards 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 predicting future performance as evaluation based on value-added.

We have a lot to learn about how to improve the reliability of value-added indicators 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 of teacher effectiveness. But when it is compared to performance assessment in other fields or to evaluations of teachers based on other sources of information, it becomes obvious that even a fuzzy signal of teacher effectiveness, if it is the best available signal, can be a vast improvement over no signal.

Teachers differ dramatically in their performance, with large consequences for students. Staffing policies that ignore this reality 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 teacher development, 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 whom they are assigned than any other source of information.

The authors make up the Brookings Brown Center Task Group on Teacher Quality. Steven Glazerman is a senior fellow at Mathematica Policy Research. Dan Goldhaber is the director of the Center for Education Data & Research at the University of Washington. Susanna Loeb is a professor of education and the director of the Institute for Research on Education Policy and Practice at Stanford University. Stephen W. Raudenbush is the Lewis-Sebring distinguished service professor in the Department of Sociology and the chair of the Committee on Education at the University of Chicago. Douglas Staiger is the John French professor of economics at Dartmouth College. Grover Whitehurst is the Herman and George R. Brown chair, a senior fellow, and the director of the Brown Center on Education Policy at The Brookings Institution.

Vol. 30, Issue 15