Problems with the Use of Student Test Scores to Evaluate Teachers

Executive Summary

Every classroom should have a well-educated, professional teacher, and school systems should recruit, prepare, and retain teachers who are qualified to do the job. Yet in practice, American public schools generally do a poor job of systematically developing and evaluating teachers. Many policy makers have recently come to believe that this failure can be remedied by calculating the improvement in students' scores on standardized tests in mathematics and reading, and then relying heavily on these calculations to evaluate, reward, and remove the teachers of these tested students.

While there are good reasons for concern about the current system of teacher evaluation, there are also good reasons to be concerned about claims that measuring teachers' effectiveness largely by student test scores will lead to improved student achievement. If new laws or policies specifically require that teachers be fired if their students' test scores do not rise by a certain amount, then more teachers might well be terminated than is now the case. But there is not strong evidence to indicate either that the departing teachers would actually be the weakest teachers, or that the departing teachers would be replaced by more effective ones. There is also little or no evidence for the claim that teachers will be more motivated to improve student learning if teachers are evaluated or monetarily rewarded for student test score gains.

A review of the technical evidence leads us to conclude that, although standardized test scores of students are one piece of information for school leaders to use to make judgments about teacher effectiveness, such scores should be only a part of an overall comprehensive evaluation. Some states are now considering plans that would give as much as 50% of the weight in teacher evaluation and compensation decisions to scores on existing tests of basic skills in math and reading. Based on the evidence, we consider this unwise. Any sound evaluation will necessarily involve a balancing of many factors that provide a more accurate view of what teachers in fact do in the classroom and how that contributes to student learning.

Evidence About the Use of Test Scores to Evaluate Teachers

Recent statistical advances have made it possible to look at student achievement gains after adjusting for some student and school characteristics. These approaches that measure growth using "value-added modeling" (VAM) are fairer comparisons of teachers than judgments based on their students' test scores at a single point in time or comparisons of student cohorts that involve different students at two points in time. VAM methods have also contributed to stronger analyses of school progress, program influences, and the validity of evaluation methods than were previously possible.

Nonetheless, there is broad agreement among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions, even when the most sophisticated statistical applications such as value-added modeling are employed.

For a variety of reasons, analyses of VAM results have led researchers to doubt whether the methodology can accurately identify more and less effective teachers. VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Another found that teachers' effectiveness ratings in one year could only predict from 4% to 16% of the variation in such ratings in the following year. Thus, a teacher who appears to be very ineffective in one year might have a dramatically different result the following year. The same dramatic fluctuations were found for teachers ranked at the bottom in the first year of analysis. This runs counter to most people's notions that the true quality of a teacher is likely to change very little over time and raises questions about whether what is measured is largely a "teacher effect" or the effect of a wide variety of other factors.

A study designed to test this question used VAM methods to assign effects to teachers after controlling for other factors, but applied the model backwards to see if credible results were obtained. Surprisingly, it found that students' fifth grade teachers were good predictors of their *fourth* grade test scores. Inasmuch as a student's later fifth grade teacher cannot possibly have influenced that student's fourth grade performance, this curious result can only mean that VAM results are based on factors other than teachers' actual effectiveness.

VAM's instability can result from differences in the characteristics of students assigned to particular teachers in a particular year, from small samples of students (made even less representative in schools serving disadvantaged students by high rates of student mobility), from other influences on student learning both inside and outside school, and from tests that are poorly lined up with the curriculum teachers are expected to cover, or that do not measure the full range of achievement of students in the class.

For these and other reasons, the research community has cautioned against the heavy reliance on test scores, even when sophisticated VAM methods are used, for high stakes decisions such as pay, evaluation, or tenure. For instance, the Board on Testing and Assessment of the National Research Council of the National Academy of Sciences stated,

…*VAM estimates of teacher effectiveness should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.*

A review of VAM research from the Educational Testing Service's Policy Information Center concluded,

VAM results should not serve as the sole or principal basis for making consequential decisions about teachers. There are many pitfalls to making causal attributions of teacher effectiveness on the basis of the kinds of data available from typical school districts. We still lack sufficient understanding of how seriously the different technical problems threaten the validity of such interpretations.

And RAND Corporation researchers reported that,

The estimates from VAM modeling of achievement will often be too imprecise to support some of the desired inferences…

and that,

The research base is currently insufficient to support the use of VAM for high-stakes decisions about individual teachers or schools.

Factors that Influence Student Test Score Gains Attributed to Individual Teachers

A number of factors have been found to have strong influences on student learning gains, aside from the teachers to whom their scores would be attached. These include the influences of students' other teachers—both previous teachers and, in secondary schools, current teachers of other subjects—as well as tutors or instructional specialists, who have been found often to have very large influences on achievement gains. These factors also include school conditions—such as the quality of curriculum materials, specialist or tutoring supports, class size, and other factors that affect learning. Schools that have adopted pull-out, team teaching, or block scheduling practices will only inaccurately be able to isolate individual teacher "effects" for evaluation, pay, or disciplinary purposes.

Student test score gains are also strongly influenced by school attendance and a variety of out-of-school learning experiences at home, with peers, at museums and libraries, in summer programs, on-line, and in the community. Well educated and supportive parents can help their children with homework and secure a wide variety of other advantages for them. Other children have parents who, for a variety of reasons, are unable to support their learning academically.
Student test score gains are also influenced by family resources, student health, family mobility, and the influence of neighborhood peers and of classmates who may be relatively more advantaged or disadvantaged.

Teachers' value-added evaluations in low-income communities can be further distorted by the summer learning loss their students experience between the time they are tested in the spring and the time they return to school in the fall. Research shows that summer gains and losses are quite substantial. A research summary concludes that while students overall lose an average of about one month in reading achievement over the summer, lower-income students lose significantly more, and middle-income students may actually gain in reading proficiency over the summer, creating a widening achievement gap. Indeed, researchers have found that three-fourths of schools identified as being in the bottom 20% of all schools, based on the scores of students during the school year, would not be so identified if differences in learning outside of school were taken into account. Similar conclusions apply to the bottom 5% of all schools.

For these and other reasons, even when methods are used to adjust statistically for student demographic factors and school differences, teachers have been found to receive lower "effectiveness" scores when they teach new English learners, special education students, and low-income students than when they teach more affluent and educationally advantaged students. The nonrandom assignment of students to classrooms and schools—and the wide variation in

students' experiences at home and at school—mean that teachers cannot be accurately judged against one another by their students' test scores, even when efforts are made to control for student characteristics in statistical models.

Recognizing the technical and practical limitations of what test scores can accurately reflect, we conclude that changes in test scores should be used only as a modest part of a broader set of evidence about teacher practice.

The Potential Consequences of the Inappropriate Use of Test-Based Teacher Evaluation

Besides concerns about statistical methodology, other practical and policy considerations weigh against heavy reliance on student test scores to evaluate teachers. Research shows that an excessive focus on basic math and reading scores can lead to narrowing and over-simplifying the curriculum to only the subjects and formats that are tested, reducing the attention to science, history, the arts, civics, and foreign language, as well as to writing, research, and more complex problem-solving tasks.

Tying teacher evaluation and sanctions to test score results can discourage teachers from wanting to work in schools with the neediest students, while the large, unpredictable variation in the results and their perceived unfairness can undermine teacher morale. Surveys have found that teacher attrition and demoralization have been associated with test based accountability efforts, particularly in high-need schools.

Individual teacher rewards based on comparative student test results can also create disincentives for teacher collaboration. Better schools are collaborative institutions where teachers work across classroom and grade-level boundaries toward the common goal of educating all children to their maximum potential. A school will be more effective if its teachers are more knowledgeable about all students and can coordinate efforts to meet students' needs.

Some other approaches, with less reliance on test scores, have been found to improve teachers' practice while identifying differences in teachers' effectiveness. They use systematic observation protocols with well-developed, research-based criteria to examine teaching, including observations or videotapes of classroom practice, teacher interviews, and artifacts such as lesson plans, assignments, and samples of student work. Quite often, these approaches incorporate several ways of looking at student learning over time in relation to a teacher's instruction.

Evaluation by competent supervisors and peers, employing such approaches, should form the foundation of teacher evaluation systems, with a supplemental role played by multiple measures of student learning gains that, where appropriate, could include test scores. Some districts have found ways to identify, improve, and as necessary, dismiss teachers using strategies like peer assistance and evaluation that offer intensive mentoring and review panels.

These and other approaches should be the focus of experimentation by states and districts. Adopting an invalid teacher evaluation system and tying it to rewards and sanctions is likely to lead to inaccurate personnel decisions and to demoralize teachers, causing talented teachers to avoid high-needs students and schools, or to leave the profession entirely, and discouraging

potentially effective teachers from entering it. Legislatures should not mandate a test-based approach to teacher evaluation that is unproven and likely to harm not only teachers, but also the children they instruct.

Economic Policy Institute. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper No. 278). Washington, DC: Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A.