Jane L. David

Using Value-Added Measures to Evaluate Teachers

Even the staunchest advocates of performance-based pay don't think it's fair to judge teachers' effectiveness solely on the basis of end-of-year test scores, without regard to where the teachers' students started at the beginning of the year. Can value-added measures, which show students' growth from one year to the next, solve this problem?

What's the Idea?

The claim for value-added measures is that they capture how much students learn during the school year, thereby putting teachers on a more level playing field as they aim for tenure or additional pay.

What's the Reality?

End-of-year test scores do not show how much students learned that year in that class, so measures that take into account where students started are surely an improvement. However, such measures of growth are only a starting point. Making judgments about individual teachers requires sophisticated analyses to sort out how much growth is probably caused by the teacher and how much is caused by other factors. For example, students who are frequently absent tend to have lower scores regardless of the quality of their teacher, so it is vital to take into account how many days students are present. Thus, to be fair and to provide trustworthy estimates of teacher effectiveness, value-added measures require complicated formulas that take into account as many influences on student achievement as possible.

What's the Research?

A growing number of researchers are studying whether value-added measures can do a good job of measuring the contribution of teachers to test score growth. Here I summarize a handful of analyses that shed light on two questions.

How Fair Are Value-Added Measures?

The trustworthiness of a value-added measure depends on how it is defined and calculated. Koretz (2008) argues that measuring the value added by the teacher requires knowing not only how much students have learned in a given year, but also the rates at which those particular students learn. Students reading well above grade level, for example, would be expected to learn faster than struggling readers. Value-added measures should take these differences into account.

Rothstein (2008) worries that test score gains are biased because students are not randomly assigned to teachers. For example, comparing teachers whose classrooms are treated as dumping grounds for troubled students with teachers whose classrooms contain the best-behaved students will favor the latter.

RAND researchers examined whether giving students different tests would lead to different conclusions about teacher effectiveness (Lockwood et al., 2006). They calculated value-added ratings of middle school teachers in a large school district on the basis of their students' end-of-year scores from one year to the next on two different math subtests. They found large differences in teachers' apparent effectiveness depending on which subtest was used. The researchers concluded that if judgments about teacher effectiveness vary simply on the basis of the test selected, administrators should use caution in interpreting the meaning of results from value-added measures.

Researchers have also identified other threats to the trustworthiness of value-added measures. Goldhaber and Hansen (2008) looked at the stability of such measures over time: Do

value-added analyses identify the same teachers as effective every year? Using a large data set from North Carolina, they found that estimates of teacher effectiveness were not the same across years in reading or math. Other researchers (for example, Koretz, 2008) question whether it is even possible to compare gains from one year to the next using tests that do not measure the same content.

Are Value-Added Measures More Accurate Than Traditional Evaluations?

Traditional methods for evaluating teacher effectiveness have their own problems—for example, infrequent or poor classroom observations or administrator bias. In fact, the persistently subjective nature of these more traditional evaluations is what fuels the current enthusiasm among policymakers for basing teacher evaluation on "objective" test scores.

Do value-added measures do a better job of judging teacher effectiveness than traditional teacher evaluations do? Researchers have looked at this question by comparing results from the two approaches.

When Jacob and Lefgren (2008) looked at 201 teachers in 2nd through 6th grade, they found a strong relationship between principals' evaluations and value-added ratings (based on student math and reading scores) of the same teachers. The researchers then asked which method did a better job of predicting how the teachers' future classes would score. They found that either method was fairly accurate in predicting which teachers would be in the top and bottom 20 percent the following year in terms of their students' test scores. Although value-added measures did a slightly better job of predicting future test scores, adding principal ratings increased the accuracy of these predictions.

Studies of teacher evaluation systems in Cincinnati, Ohio, and Washoe County, Nevada, also found that value-added measures and well-done evaluations based on principal observations produced similar results (Milanowski, Kimball, & White, 2004).

What's One to Do?

From the federal government to foundations, the pressure is on to use student test score gains to evaluate teachers. Yet doing so in a credible and fair way is a complex and expensive undertaking with no guarantee that intended improvements in teaching and learning will result. What's more, it is not clear that value-added measures yield better information than more traditional teacher evaluation practices do.

To protect teachers from erroneous and harmful judgments, we need multiple measures.

The complexity and uncertainty of measuring student achievement growth and deciding how much responsibility for gains to attribute to the teacher argue against using such measures for high-stakes decisions about individuals. To protect teachers from erroneous and harmful judgments, a consensus is emerging that we need multiple measures that tap evidence of good teaching practices as well as a variety of student outcomes, including but not limited to standardized test score gains. According to a recent study (Coggshall, Ott, & Lasagna, 2010), most teachers support such a multiple-measures approach.

Investing in expensive data analysis systems may not be as important as investing in ways of measuring teacher effectiveness that can identify the specific supports teachers need to improve their practice. EL

References

Coggshall, J. G., Ott. A., & Lasagna, M. (2010). *Convergence and contradictions in teachers' perceptions of policy reform ideas* (Retaining Teacher Talent, Report No. 2). Naperville, IL: Learning Point Associates and New York: Public Agenda. Available: www.learningpt.org/expertise/educatorquality/genY/CommunicatingReform/index.php

Goldhaber, D., & Hansen, M. (2008). *Is it just a bad class? Assessing the stability of measured teacher performance.* (Working paper #2008-5). Seattle: Center on Reinventing Public Education, University of Washington. Available: www.crpe.org/cs/crpe/view/csr_pubs/249

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136. Available: http://ideas.repec.org/a/ucp/jlabec/v26y2008p101-136.html

Koretz, D. (2008, Fall). A measured approach: Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator,* 18–27, 39.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V-N., & Martinez, F. (2006). *The sensitivity of value-added teacher effect estimates to different mathematics achievement measures.* Santa Monica, CA: RAND Corporation. Available: www.rand.org/pubs/reports/2009/RAND_RP1269.pdf

Milanowski, A., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites.* Madison: Consortium for Policy Research in Education, University of Wisconsin. Available: http://cpre.wceruw.org/papers/3site_long_TE_SA_AERA04TE.pdf

Rothstein, J. (2008). *Student sorting and bias in value added estimation: Selection on observables and unobservables.* (Working Paper No. 170). Princeton, NJ: Princeton University and Cambridge, MA: National Bureau of Economic Research.

Jane L. David is Director of the Bay Area Research Group, Palo Alto, California; jld@bayarearesearch.org.