ACT Research & Policy

Issue Brief

September 2014

Evidence suggests that value-added measures of teacher effectiveness can be a valuable tool to improve teacher evaluation, identify teachers on the extremes of effectiveness, and identify factors that improve student performance in the classroom.

Recent Validity Evidence for Value-Added Measures of Teacher Performance

Richard Buddin and Michelle Croft

Value-added measures are becoming a common component in teacher evaluations. By the 2016–2017 school year, most if not all states will have implemented a teacher evaluation system that includes the use of value-added measures.[1] This shift to include student achievement data in teacher evaluations is not without its critics. However, many of the criticisms prevalent today were levied when value-added measures were first gaining popularity in the mid-2000s. Since that time a number of rigorous studies have addressed these criticisms and provided validity evidence to support the use of value-added measures as a component of teacher evaluation.

The criticisms have likely persisted despite the evidence supporting value-added measures because the value-added measures are such a departure from traditional teacher evaluation systems. Under the traditional system, evaluations are based on short classroom observations by school principals or other school administrative personnel where nearly all teachers receive the highest ratings.[2] Basing teachers' evaluations on their students' academic performance is a new use of student test scores. Also, the value-added models are more technical (though less subjective) than classroom observations, and they are not always communicated in a way that parents, teachers, or school administrators can easily understand.[3]

In this report, we (1) explain what we mean by value-added measures, (2) identify the common criticisms of value-added measures and the research evidence that addresses those criticisms, and (3) provide validity evidence for the use of value-added measures in teacher evaluations.

What Value-Added Measures Are

Over the past decade, numerous studies have used longitudinal student-level data to estimate the contribution of teachers to student learning.[4] The methods these studies have relied upon, called value-added methods, isolate teacher contributions to student outcomes by estimating the effects of teachers on student achievement conditional on prior-year test scores and student-level measures of student demographics and background. The value-added approach relies on teacher "output" as measured by improvements in student test scores. This approach is a sharp departure from orthodox measures of teacher quality that have relied on teacher preparation and training (e.g., education level, experience, or subject matter knowledge) and occasional classroom observations by a school administrator.

Researchers of value-added measures typically find wide variability in teacher effects, suggesting that some teachers may be much more effective than others at improving student achievement. Some findings are common across most studies.

- **Experience.** New teachers are typically less effective than others, but teacher effects vary little with experience after the first year or two of teaching.[5]
- **Advanced degrees.** Teachers with master's degrees have similar effects to teachers with only bachelor's degrees.[6]
- **Certification.** Teachers with alternative certification are often just as effective at improving test scores as teachers certified through traditional programs.[7]

- **Distribution of teacher value-added scores.** Teachers with high value-added results are widely distributed across schools and not concentrated in a few schools. After controlling for prior achievement, teachers in intercity schools often perform as well or better than their counterparts in more wealthy suburban schools. Teacher value-added scores vary more within schools than across schools.[8]

Common Criticisms of Value-Added Measures and Research Evidence Supporting Their Use

Most criticism of value-added approaches has focused on concerns about ranking individual teacher performance. Several studies have raised concerns about value-added assessment, suggesting that value-added measures may provide either little information regarding the effectiveness of a teacher or misleading information about the effect a teacher may have on his or her students. However, a substantial amount of research provides credible validity evidence to support the appropriate use of value-added measures. This section discusses the criticisms of value-added measures and reviews research evidence that addresses these criticisms.

Criticism 1: Achievement tests aren't completely accurate, so they should not be used to evaluate teachers.

Student achievement tests are incomplete measures of student knowledge, and even the best tests measure achievement with some error.[9] This inherent measurement error means that any aggregation of a teacher's students' scores will misclassify some teachers as effective or ineffective.

Research Evidence: Despite the presence of measurement error in state tests, researchers have found that teacher value-added estimates do predict the achievement of students assigned to a teacher.

Education, like engineering and health, invariably relies on measures that are subject to some error. The efficacy of a new measure should be based on whether the new measure provides a better prediction of the intended outcome than the current measures. Ideally, the new measures would be perfect or very precise, but decision makers cannot wait for the perfect measure.[10] Educators have long relied on achievement tests to identify whether students need academic assistance, are ready for advanced coursework, or are ready for college. The key question for policymakers is whether the use of student achievement as an element of teacher evaluation will provide useful information for improving instruction or the quality of the teacher workforce.

Critics have argued that measurement error means that value-added estimates of teacher effectiveness would vary substantially from one test to another. Researchers have found similar value-added results when the same students take different tests. For instance, researchers as part of the Measures of Effective Teaching (MET) Project compared teachers' value-added estimates from state tests with those from separate math and English assessments that measured higher-order skills.[11] Some critics have argued that teacher effects derived from state tests are misleading because these scores reflect "teaching to the test" and not a deeper level of student learning.[12] However, the researchers found that student scores were highly correlated on the two sets of tests, so teachers that had high value-added scores on the state test in math or English were likely to have high value-added scores on the higher-order tests as well.[13]

In addition to measurement error, there may be test ceiling effects where the student is unable to demonstrate growth due to the difficulty level of the test, which would affect value-added measures. However, with the exception of some state minimum competency tests, researchers have not found ceiling effects that influence value-added estimates.[14]

Criticism 2: Value-added measures don't take into account all of the potential background variables that may affect a student's test score, including teacher assignment.

A second criticism of value-added methods is that most studies rely on district or state administrative data and have few controls for the mix of students assigned to an individual teacher.[15] For example, these control variables are often limited to gender, race/ethnicity, free/reduced lunch eligibility, English learner status, and special education status. With limited controls, researchers are unable to adjust estimates for the possible sorting of students into classrooms.[16] If students are nonrandomly assigned to classrooms, then teacher value-added scores may reflect nuances of the sorting mechanism instead of differences in actual teacher effectiveness. If some teachers are assigned "better" students than others, then they have an unfair advantage in the teacher ranking. Since value-added measures control for prior achievement, "better" in this context is not simply high-achieving students, but rather students with more potential for improvement in a given year.

Research Evidence: In practice, the contextual controls and student sorting have not created large distortions in value-added estimates.

Chetty, Freedman, and Rockoff examined potential biases in estimated teacher effects due to the sorting of students into classrooms.[17] For example, administrative data have weak measures of family socioeconomic status (SES). If high-SES

students were disproportionately concentrated with some teachers, then these teachers might appear more effective than others simply because of the selection of students into their classrooms. Chetty, Freedman, and Rockoff tested this sorting bias by comparing teacher effects from traditional administrative data versus data that included richer controls for family SES. The family characteristics included parental marital status, family income, mother's age at student's birth, and indicators for parental contributions to a 401(k) and home ownership. The researchers found that the absence of the family-level variables (e.g., marital status and income) in traditional estimates of teacher effects had little effect on those estimates. They argue that the bias in the teacher effects is small because the effects of these family characteristics are implicitly included in the traditional models through controls for lagged test scores.

Chetty, Freedman, and Rockoff also examined whether estimated teacher effects were consistent with changes in grade-level test scores as teachers moved from school to school. If the value-added methodology accurately captures persistent differences in teacher effectiveness, then the movement of a high-quality teacher from one school to another should have a predictable effect on achievement at both the old and new school. For example, when a high-quality fourth grade teacher moves to a new school, the grade-level gains should decrease at the school the teacher left and increase at the school the teacher enters. Indeed, student test scores moved as predicted when teachers moved from school to school, providing further evidence that estimated teacher effects represent a persistent and real underlying difference in teacher effectiveness.

The landmark MET Project provides a substantial counter to the criticism of nonrandom student assignment to teachers.[18] This massive study was conducted across six large metropolitan school districts. Teacher effectiveness was measured both through multiple classroom evaluations by trained observers and through value-added techniques. Teacher value-added scores were measured for an initial period and compared with estimates following the random assignment of students to teachers. Students were randomly assigned to teachers, so observed teacher effectiveness was not confounded by the types of students assigned to different teachers.

The researchers found that student sorting had little effect on value-added estimates of teacher effectiveness. Teacher rankings before and after random assignment were highly correlated with one another. This evidence suggests that student-level controls in value-added models may be adequate controls for differences in students assigned to individual teachers.

Criticism 3: Value-added measures aren't stable from year to year.

Another criticism of value-added measures is that value-added estimates may vary from year to year as a teacher's classes or effectiveness changes over time.[19] If so, then value-added estimates would provide limited insight into which teachers or teacher practices were most effective.

Research Evidence: Value-added measures are unstable from year to year, but they are stable over multiple years and classes.

Several empirical studies show that value-added scores are relatively unstable from year to year, but the studies also find persistent, stable teacher effects when teachers are observed across multiple years and classes.[20] The instability of year-to-year estimates reflects measurement error in the tests as well as the small number of students taught by teachers in a given year. For example, elementary school classes often contain only 20 to 30 students, so teacher effects may be unduly affected by test results for a small number of students. The evidence suggests that this instability is sharply reduced if the teacher effects are based on even two or three years of data.

School officials, teachers, and parents may prefer a "real time" metric on how teachers are performing each year. While value-added methods provide a cumulative measure of teacher effectiveness, the research evidence suggests that these methods should not be applied to an individual class. In a similar vein, the classroom observation method gives a subjective estimate of teacher effectiveness during the class that was observed and may not reflect a teacher's effectiveness in other classes or on other days. Neither measure provides a perfect indication of teaching effectiveness in all circumstances, but the key policy question is whether value-added measures, when correctly applied, are an appropriate tool for improving teacher evaluation.

Criticism 4: Value-added measures aren't related to other teacher quality measures.

Haertel criticizes value-added methods because they provide no direct indication of why some teachers are more effective than others or how individual teachers could improve.[21] In contrast, classroom observation and teacher pedagogy approaches provide better hands-on recommendations for improving instruction.

Research Evidence: Value-added measures are highly correlated with teaching practices.

While value-added studies provide no information on classroom practices per se, a few recent studies have shown linkages between value-added estimates and teaching

practices. This linkage suggests that teachers with low value-added scores may be able to improve their effectiveness and value-added scores by implementing better teaching practices.

Kane, Taylor, Tyler, and Wooten[22] combined data on teacher value-added scores in Cincinnati with multiple classroom evaluations of each teacher by trained evaluators.[23] The study found that value-added measures and classroom observations were highly correlated and that improvements in classroom practices were likely to improve student achievement growth.

Grossman, Loeb, Cohen, and Wyckoff examined the relationship between instructional practices and value-added assessments for middle school English Language Arts teachers.[24] The study relied on trained evaluators observing instructional practices using an observational protocol. They found that high value-added teachers employ much more effective instructional practices than low value-added teachers.

The MET Project included detailed classroom evaluations by multiple trained evaluators as well as value-added estimation.[25] The results showed substantial variability of teachers by different evaluators even with detailed evaluation protocols. The classroom evaluations under various protocols were highly correlated with teacher value-added scores on both state and higher-order skills tests. This study reinforces the notion that value-added estimates reflect underlying differences in instructional practice, and the estimates are consistent with recently developed instructional protocols.

The MET Project concluded that teacher evaluation should include value-added estimates, multiple and detailed classroom observations, and student survey information. Value-added estimates provide a cost-efficient method for differentiating less from more effective teachers, and multiple detailed classroom observations can provide teachers with meaningful feedback to improve their instructional practices. Combining the measures produces a more robust teacher evaluation.

Validity Evidence to Support Use of Value-Added Measures

In addition to other findings, recent predictive evidence from value-added assessment provides further support for the validity of the approach. Chetty, Freedman, and Rockoff studied twenty years of data on students and teachers in third through eighth grade in a large metropolitan school district.[26] The researchers constructed value-added estimates of teacher effects and tracked the long-term effects of teachers on the adult outcomes of their students. They found that students with primary school teachers who had high value-added scores were more likely to attend college, earn high wages as adults, live in higher SES neighborhoods, and have higher savings rates. These persistent effects of teachers with high value-added scores on students provide evidence in support of the validity of value-added measures.

Glazerman, Protik, Bruch, and Max provided additional evidence on the persistence of teacher effects as teachers move from school to school.[27] Their study investigated the use of financial incentives to encourage teachers with high value-added scores (top 20%) to volunteer for an assignment in low-achieving schools. Vacant teaching positions were randomly filled by either a high value-added teacher with a $20,000 incentive (the treatment group) or by another teacher through normal hiring practices. The high value-added teachers had positive effects on elementary test scores in the low-achieving schools relative to the group of control teachers.[28]

Finally, Dee and Wyckoff analyzed the effects of a Washington, DC, teacher evaluation system that was based on a combination of structural classroom evaluations and value-added assessments.[29] They found that the evaluation system encouraged low-performing teachers to voluntarily leave district positions, and those who remained made large student achievement gains in their classrooms. In addition, financial incentives for high-performing teachers were associated with high classroom achievement gains. This evidence suggests that value-added measures may be an important component of teacher evaluation that could substantially improve student achievement.

Summary

Teachers are an essential part of student learning. They have long-term, substantial, and substantive impacts on students. "Competent teachers are a critical, if not the most important, component of success of a child's in-school educational experience [and ineffective] teachers substantially undermine the ability of that child to succeed in school," wrote Judge Treu when ruling that California's teacher tenure and dismissal statutes violated the California state constitution in *Vergara v. California*.[30]

There is a need to ensure that students receive an appropriate education and are not assigned to highly ineffective teachers. To ensure this, we must first be able to accurately identify a highly ineffective teacher with some measure of objectivity, and despite criticisms of value-added measures, when used in conjunction with other methods to give teachers meaningful feedback to improve instruction, they provide an effective and efficient way of doing so. However, more efforts are needed to ensure that educators understand how value-added measures work and the research behind the measures to increase transparency in the evaluation process.

Value-added measures, like all other measures, are imperfect indications of teacher effectiveness, but the research evidence suggests that they are probably an improvement over the current evaluation system. Performance, like student learning itself, is always measured with some degree of error, but this imperfection should not deter policymakers from carefully considering student achievement outcomes in making decisions. Recent validity evidence on value-added measures suggests that these methods can be a valuable tool for improving teacher evaluation, identifying teachers on the extremes of the effectiveness distribution, and identifying factors that improve student performance in the classroom. ■

Notes

1 US Department of Education, "ESEA Flexibility State-by-State Implementation Timeline Chart," June 14, 2013, http://www2.ed.gov/policy/elsec/guid/esea-flexibility/eseaflexstchart614.doc.

2 Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling, *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (Brooklyn, NY: The New Teacher Project, 2009), http://widgeteffect.org.

3 For example, in an interview with the *Tampa Bay Times*, Florida House Speaker Will Weatherford told the paper's editorial board "I think the concept works. . . . But it's such a complex thing, I frankly couldn't even explain it to you." Lisa Gartner and Cara Fitzpatrick, "Confused by Florida's Teacher Scoring? So are Top Teachers," *Tampa Bay Times*, March 1, 2014, http://www.tampabay.com/news/education/k12/confused-by-floridas-teacher-performance-scores-so-are-award-winning/2168062.

4 Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73, no. 2 (2005): 417–458, doi:10.1111/j.1468-0262.2005.00584.x; Robert Gordon, Thomas Kane, and Douglas O. Staiger, *Identifying Effective Teachers Using Performance on the Job* (Brookings Institution Discussion Paper 2006-1) (Washington, DC: The Brookings Institution, 2006); Douglas Harris and Tim R. Sass, "Teacher Training, Teacher Quality and Student Achievement," *Journal of Public Economics* 95, no. 7–8 (August 2011): 798–812, doi:10.1016/j.jpubeco.2010.11.009; Daniel Aaronson, Lisa Barrow, and William Sander, "Teachers and Student Achievement in Chicago Public High Schools," *Journal of Labor Economics* 24, no. 1 (2007): 95–135, http://www.jstor.org/stable/10.1086/508733; Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "How and Why Do Teacher Credentials Matter for Student Achievement?" (NBER Working Paper No. 12828) (Cambridge, MA: National Bureau of Economic Research, 2007); Cory Koedel and Julian Betts, "Re-examining the Role of Teacher Quality in the Educational Production Function" (working paper, San Diego, CA: University of California, San Diego, 2007); Brian A. Jacob and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics* 26, no. 1 (2008): 101–136, http://www.jstor.org/stable/10.1086/522974; Thomas J. Kane, Jonah E. Rockoff, and Douglas O. Staiger, "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27 (2008): 615–631, doi:10.1016/j.econedurev.2007.05.005; Thomas J. Kane and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," (NBER Working Paper No. 14607) (Cambridge, MA: National Bureau of Economic Research, 2008); Richard Buddin and Gema Zamarro, "Teacher Qualifications and Student Achievement in Urban Elementary Schools," *Journal of Urban Economics* 66 (2009): 103–115, doi:10.1016/j.jue.2009.05.001; Daniel F. McCaffrey, Tim R. Sass, J.R. Lockwood, and Kata Mihaly, "The Intertemporal Variability of Teacher Effects Estimates," *Education Finance and Policy* 4, no. 4 (2009): 572–606, doi:10.1162/edfp.2009.4.4.572; and Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas Staiger, *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment* (Seattle, WA: Bill and Melinda Gates Foundation, 2013).

5 Harris and Sass, *The Effects of Teacher Training*; Kane, Rockoff, and Staiger, "What Does Certification Tell Us"; and Buddin and Zamarro, "Teacher Qualifications and Student Achievement."

6 Koedel and Betts, "Re-examining the Role of Teacher Quality"; Aaronson, Barrow, and Sander, "Teachers and Student Achievement"; and Buddin and Zamarro, "Teacher Qualifications and Student Achievement."

7 Kane, Rockoff, and Staiger, "What Does Certification Tell Us"; and Tim Sass, *Certification Requirements and Teacher Quality: A Comparison of Alternative Routes to Teaching* (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, American Institutes for Research, 2011).

8 Rivkin, Hanushek, and Kain, "Teachers, Schools, and Academic Achievement"; and Buddin and Zamarro, "Teacher Qualifications and Student Achievement."

9 Edward H. Haertel, *Reliability and Validity of Inferences about Teachers Based on Student Test Scores* (Princeton, NJ: Educational Testing Service, 2013); and Eva L. Baker, Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard, *Problems With the Use of Student Test Scores to Evaluate Teachers* (Washington, DC: Economic Policy Institute, 2010).

10 Steven Glazerman, Dan Goldhaber, Susanna Loeb, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst, "Value-Added: It's Not Perfect, But It Makes Sense," *Education Week*, December 15, 2010.

11 MET Project, *Ensuring Fair and Reliable Measurs of Effective Teaching* (MET Project Policy and Practice Brief) (Seattle, WA: Bill & Melinda Gates Foundation, 2013), http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

12 Baker et al., *Problems with the Use of Student Test Scores*.

13 The tests of higher-order skills included the SAT 9 Open-Ended Reading Assessment (SAT 9 OE) and the Balanced Assessment in Mathematics (BAM). MET Project, *Ensuring Fair and Reliable Measures of Effective Teaching*.

14 Koedel and Betts, "Re-examining the Role of Teacher Quality."

15 Haertel, *Reliability and Validity of Inferences*; and Baker et al., *Problems with the Use of Student Test Scores*.

16 Jesse Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4, no. 4 (2009): 537–571, doi:10.1162/edfp.2009.4.4.537; and Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125, no. 1 (2009): 175–214, doi:10.1162/qjec.2010.125.1.175.

17 Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates" (NBER Working Paper No. 19423) (Cambridge, MA: National Bureau of Economic Research, 2013), doi:10.3386/w19423.

18 MET Project, *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* (MET Project Research Paper, Seattle, WA: Bill & Melinda Gates Foundation, 2012), http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf; and Kane et al., *Have We Identified Effective Teachers?*

19 Baker et al., *Problems with the Use of Student Test Scores*.

20 Koedel and Betts, "Re-examining the Role of Teacher Quality"; Richard Buddin, Daniel F. McCaffrey, Sheila Nataraj Kirby, and Nailing Xia, *Merit Pay for Florida: Design and Implementation Issues* (WR-508-FEA) (Santa Monica, CA: RAND, 2007); and McCaffrey et al., "The Intertemporal Variability of Teacher Effects."

21 Haertel, *Reliability and Validity of Inferences*.

22 Thomas J. Kane, Eric S. Taylor, John H. Tyler, and Amy L. Wooten, "Identifying Effective Classroom Practices Using Student Achievement Data," *Journal of Human Resources* 46, no. 3 (2010): 587–613.

23 The evaluations were based on the framework developed by Charlotte Danielson in *Enhancing Professional Practice: A Framework for Teaching*, 2nd ed. (Washington, DC: Association for Supervision and Curriculum Development, 2007).

24 Pam Grossman, Susanna Loeb, Julia Cohen, and James Wyckoff, "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores," *American Journal of Education* 119, no. 3 (2013): 445–470.

25 MET Project, *Gathering Feedback*; and MET Project, *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching* (Seattle, WA: Bill & Melinda Gates Foundation, 2013), http://metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf.

26 Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," (NBER Working Paper No. 19424) (Cambridge, MA: National Bureau of Economic Research, 2013).

27 Steven Glazerman, Ali Protik, Bing-ru TehJulie Bruch, and Jeffrey Max, *Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment* (Washington, DC: US Department of Education, 2013).

28 The study did not find significant differences in middle school student achievement between the treatment and control groups. The authors argue that the insignificance of the middle school results may reflect small sample sizes in the middle school analysis or district-specific issues in the study districts.

29 Thomas Dee and James Wyckoff, "Incentive, Selection, and Teacher Performance: Evidence from IMPACT" (NBER Working Paper No. 19529) (Cambridge, MA: National Bureau of Economic Research, 2013), doi:10.3386/w19529.

30 Vergara v. State of California, Tentative Decision, No. BC484642 (Cal. Super. Ct.).