Can Value Added Add Value to Teacher Evaluation?

Linda Darling-Hammond[1]

Keywords: accountability; descriptive analysis; education policy; teacher assessment; teacher evaluation

The five thoughtful papers included in this issue of *Educational Researcher* (*ER*) raise new questions about the use of value-added methods (VAMs) to estimate teachers' contributions to students' learning as part of personnel evaluation. The papers address both technical and implementation concerns, considering potential effects on teachers' behavior and on the resulting quality of teaching. In this response, I reflect on articles' findings in light of other work in this field, and I offer my own thoughts about whether and how VAMs may add value to teacher evaluation.

Value Added in Theory

The notion of using VAMs to evaluate educators and schools is intuitively appealing. The quality of teaching, most would agree, is signaled by how well students are learning. Like many other researchers who have used VAMs in studies of program and policy effects, I was initially enthusiastic about its possibilities. From my research, I am also well-aware of the historical shortcomings of many districts' teacher evaluation practices (Darling-Hammond, 2013). These have stimulated federal incentives to transform evaluation through systematic classroom observations coupled with student learning evidence.

The question of whether value-added ratings will ultimately improve or undermine teacher evaluation depends in large measure on whether VAM metrics can accurately identify individual teachers' contributions to student learning and hence offer a credible measure of teacher "effectiveness." In theory, VAMs could do so under a set of ideal conditions, if:

- student learning is well-measured by tests that reflect valuable learning and the actual achievement of individual students along a vertical scale representing the full range of possible achievement measured in equal interval units;
- students are randomly assigned to teachers within and across schools—or, conceptualized another way, the learning conditions and traits of the group of students assigned to one teacher do not vary substantially from those assigned to another; and
- individual teachers are the only contributors to students' learning over the period of time used for measuring gains.

Of course, none of these assumptions holds, and the degree of error in measuring learning gains and attributing them to a specific teacher depends on the extent to which they are violated, as well as the extent to which statistical methods can remedy these problems.

Unfortunately, in the United States, at this moment in history, the violations of these assumptions are considerable. With respect to assessment, standardized tests in the United States are criticized for their narrowness and focus on lower level skills; evidence has shown that high-stakes incentives to focus on these tests have reduced time spent teaching other important content and skills (Darling-Hammond & Adamson, 2014). Furthermore, because the No Child Left Behind Act mandated that state tests measure grade-level standards only, the tests do not include items that assess content or skills from earlier or later grade levels. As a result, these tests cannot measure the actual achievement level—or the learning gains—of the large share of students who are above or below grade level in their knowledge and skills.

As Ed Haertel (2013), chair of the National Research Council's Board on Testing and Assessment, has noted, this feature of test design "translates into bias against those teachers working with the lowest-performing or the highest-performing classes" (p. 8). This bias is indicated in many studies that find that VAM measures appear particularly inaccurate for teachers whose students who achieve below or above grade level, who are new English learners, or who have special needs (Haertel, 2013; Newton, Darling-Hammond, Haertel, & Thomas,

[1]Stanford University, Stanford, CA

Educational Researcher, Vol. 44 No. 2, pp. 132–137
DOI: 10.3102/0013189X15575346

2010) and for those who teach in tracked school settings (Harris & Anderson, 2011; Jackson, 2012).

The new tests created by the Partnership for Assessing Readiness for College and Careers (PARCC) and Smarter Balanced, the multistate consortia created to evaluate the Common Core State Standards, will not remedy this problem as they, too, have been required to measure grade-level standards. Even though they will report students' scores on a vertical scale, they will not be able to measure accurately the achievement or learning of students who started out below or above grade level.

With respect to the equivalence of student groups across classrooms, the U.S. education system is the one of most segregated and unequal in the industrialized world. The country's extraordinarily high rates of childhood poverty, homelessness, and food insecurity are not randomly distributed across communities. Wealthy enclaves are increasingly segregated by race and class from poor ones. The fact that schools and districts have extremely unequal funding, combined with the tattered safety net for children, means that teachers working in lower income communities often have fewer resources to serve concentrations of students with much greater educational, psychological, health, and social needs. Add the extensive practice of tracking to the mix, and it is clear that the assumption of equivalence among classrooms is far from reality.

Finally, we know from decades of educational research that many things matter for student achievement aside from the individual teacher a student has at a moment in time for a given subject area. A partial list includes the following:

- School factors such as class sizes, curriculum choices, instructional time, availability of specialists, tutors, books, computers, science labs, and other resources;
- prior teachers and schooling, as well as other current teachers—and the opportunities for professional learning and collaborative planning among them;
- peer culture and achievement;
- differential summer learning gains and losses;
- home factors, such as parents' ability to help with homework, food and housing security, and physical and mental support or abuse; and
- individual student needs, health, and attendance.

Given all of these influences on learning, it is not surprising that variation among teachers accounts for only a tiny share of variation in achievement, typically estimated at under 10%. The American Statistical Association (ASA, 2014), in its statement on VAMs, noted that:

> most VAM studies find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions. Ranking teachers by their VAM scores can have unintended consequences that reduce quality. (p. 2)

A few of the nonteacher factors are measured in some of the VAM models; most assume that controlling for prior test scores takes care of unmeasured influences on gains. The unmeasured variables that influence achievement gains become part of the so-called "teacher effect" that, as one statistician quipped, is really just the error term in the regression equation. I return to the question of whether these concerns can be adequately addressed at the end of this article. Suffice it to say that they pose considerable challenges to deriving accurate estimates of teacher effects, and as the ASA suggests, these challenges may have unintended negative effects on overall educational quality.

Value Added in Practice

In various ways, these five articles illustrate outcomes of the challenges identified above. Dan Goldhaber (this issue, pp. 87–95) cites a number of studies that have documented the instability of estimates from year to year, class to class, and test to test, as well as across statistical models. The degree of fluctuation can be quite large: A study examining data from five school districts found, for example, that of teachers who scored in the bottom 20% of rankings in one year, only about a quarter remained there in the following year, whereas about 50% scored in the top half. The same volatility occurred for those who scored at the top of the distribution in Year 1, about half of whom moved to the bottom half of the distribution in Year 2 (Sass, 2008).

Similarly, when researchers used a different statistical model to recalculate the value-added scores in reading and mathematics for teachers whose scores were published in the *Los Angeles Times* in 2011, they found that 40% to 55% of teachers would get noticeably different scores (Briggs & Domingue, 2011).

Many studies have found significantly different outcomes for teachers' rankings on different tests in their content area. In one such study, 20% to 30% of teachers who ranked in the top quartile of VAM ratings on state tests of basic skills ranked in the bottom half of impacts on more conceptually demanding tests of higher order skills and vice versa (Rothstein, 2011).

So which is correct? The measure in Year 1 or the one in Year 2? The rating produced by one statistical model or the rating produced by another? The metric resulting from Test 1 or the one produced by Test 2? These are not small differentials and in current high-stakes contexts can mean the difference between a teacher being rewarded with a bonus or being fired.

These examples illustrate how teachers can be misclassified. Goldhaber (this issue) notes that the effects of value added on the quality of the teaching force depend both on the extent of teacher misclassification and on how teachers react to these measures—whether good teachers are motivated to enter and stay in the profession and whether poor ones are motivated, or forced, to leave. He concludes that the jury is still out on whether the use of value added will improve or undermine the teaching force in the long run.

Dale Ballou and Matthew Springer (this issue, pp. 77–86) add to the list of concerns, noting that value-added estimates are "notoriously imprecise" and urging that users acknowledge these estimation errors. They point to the challenges with the Colorado Growth Model and its variants in states like Georgia, which particularly disadvantages teachers with small numbers of students, as well as to the Educational Value-Added Assessment System (EVAAS), which puts teachers with more students at greater risk of misclassification because of an inappropriate statistical interpretation. Both systems are used to make high-stakes decisions without acknowledging these limitations.

Error ranges can be extremely large. As Sean Corcoran (2010) documented with New York City data, after taking statistical

uncertainty into account, the "true" effectiveness of a teacher ranked in the 43rd percentile on New York City's Teacher Data Report might have a range of possible scores from the 15th to the 71st percentile, qualifying as "below average," "average," or close to "above average." Even using multiple years of data, the error range is still so large that "half of all teachers in grades four to eight cannot be statistically distinguished from 60 percent or more of all other teachers in the city." Corcoran noted, "It is unclear what this teacher or his principal can do with this information to improve instruction or raise student performance" (p. 6).

Ballou and Springer (this issue) advise responsible means for better representing and reporting the degree of error. It is unclear whether practitioners would know how to interpret this information, but it is highly likely that if educators knew how much error is associated with these metrics, their current low levels of confidence in the measures would be shaken even further.

Ballou and Springer (this issue) also take up the issue of whether value-added estimates should be annually revised, as they are in the EVAAS system, based on results of students in years after the teacher had them. The authors note that "this has confused teachers, who wonder why their value-added score keeps changing for students they had in the past." Furthermore, they add, "What will be done about the teacher whose performance during the 2013-14 school year, as calculated in the summer of 2014, was so low that she loses her job or her license, but whose revised estimate for the same year, released in the summer of 2015, places her performance above the threshold at which these sanctions would apply?" Although they acknowledge there may be sound statistical reasons for this practice (although they do not comment on how the presumed effectiveness of the later teachers is supposed to be factored out of these data), their analysis points to the profound practical problems associated with attaching high stakes to measures that are so imprecise and unstable.

Thus, it should not be surprising that the other studies in this issue point to pervasive concerns on the part of educators about VAM measures. Ellen Goldring et al. (this issue, pp. 96–104) note that, in the six urban districts where their team surveyed and interviewed hundreds of administrators, leaders identified "numerous shortcomings in the usefulness of student test score-based models," including concerns about validity, lack of transparency, and complexity. By contrast, in these districts piloting new models, the use of improved teacher observation protocols, with multiple, trained evaluators, was proving to be more helpful in the evaluation process.

The study notes that "principals across all school systems revealed major hesitations and challenges regarding the use of value-added measures for human capital decisions." They often perceived the VAM metrics as "inflated" or "deceptive" and felt pressured to make their observation ratings align with value-added measures that they saw as inaccurate. A central office administrator noted that the outcomes produced by these measures were mysterious to teachers, who could not explain why they fluctuated from year to year when their practice did not, noting "we've not successfully been able to articulate that for teachers." This comment echoes the report of a mystified teacher in Houston, who remarked:

> I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus. … What did I do differently? I have no clue. (Amrein-Beardsley & Collins, 2012, p. 15)

Jennie Jiang, Susan Sporte, and Stuart Luppescu (this issue, pp. 105–116) report similar skepticism among teachers they surveyed in Chicago. Like the administrators in the other cities, Chicago teachers felt positively about the new observation process, which had been strengthened with a new rubric and training for evaluators, but were concerned about the inclusion of the value-added scores in their evaluations. Survey responses found that most teachers felt their evaluators could accurately assess their instruction and offered useful feedback for improving teaching, and these positive perceptions increased between the first and second years of implementation.

Yet satisfaction with the overall system went down significantly. Dissatisfaction was associated with the value-added component of the rating. Sixty-five percent of teachers reported that their evaluation "relies too heavily on student growth," and 50% felt that the test data were not an accurate measure of their students' learning. Teachers voiced concerns about the narrowness of learning represented on the standardized tests, the increase in testing burdens, and perceptions that the measures are unfair to teachers working in challenging schools where "things that a teacher cannot possibly control" substantially influence how children do in school.

Susan Moore Johnson's article (this issue, pp. 117–126) reinforces this concern, citing studies that have found VAM models unable to fully account for the differences in student backgrounds and learning differences. Johnson notes that the use of VAMs may discourage teachers from working in high-need schools or with high-need students, making these classrooms and schools even harder to staff than they already are.

Researchers report such incentives operating in Houston where teachers have noted their value-added scores go down when they are assigned to teach in fourth grade where English learners are transitioned into mainstreamed classrooms, and this dip leads to dismissals. One teacher commented, "I'm scared I might lose my job if I teach in a transition grade level, because … my scores are going to drop." Another explained, "When they say nobody wants to do 4th grade—nobody wants to do 4th grade. Nobody!" (Amrein-Beardsley & Collins, 2012, p. 16).

Johnson (this issue) raises the further possibility that the use of value-added measures may inhibit the formation of the social capital that enables a school organization to become "greater than the sum of its parts" through the collegial activities that are associated with greater student learning gains than isolated teachers can produce. She worries that statisticians' efforts to parse out learning to individual teachers may cause teachers to hunker down and focus only on their own students, rather than working collegially to address student needs and solve collective problems.

This response was illustrated by a teacher addressing the proposed use of VAMs to determine teacher pay in a poor district where budget cuts had stalled salaries for years and teacher turnover reached 50%. Susan Saunders, a respected veteran, had stayed and worked tirelessly to assist the revolving door of beginning teachers. She gave the only textbooks to the new teachers and took on the special education students (comprising most of her class of 32 students), because she was able to work with them successfully.

Saunders explained that the test-based pay system would cause her, sadly, to stop taking on the special education students and sharing materials with other teachers, as she could no longer help others without hurting herself (Darling-Hammond, 2013).

Can Value Added Add Value to Teacher Evaluation?

Still, the concept of considering teachers' contributions to student learning is appealing to policymakers, and with federal incentives, it is now embedded in policy in more than 30 states. Can value-added measures be made accurate and credible? More important, can VAMs become a positive force for building a more effective and equitable system of education, rather than a disincentive for teachers to serve high-need students and to work collaboratively with each other? Equally important, given that VAM measures appear arbitrary and error-prone, is whether their use will dissuade smart, capable people from entering and staying in a profession in which they believe a considerable portion of their evaluation could be volatile and inaccurate.

If useful answers to these questions are to be found, we will need more of the close analysis of technical and practical concerns found in these articles to guide discussion of VAMs, which has been prematurely thrust into policy contexts that have made it more the subject of advocacy than of careful analysis that shapes its use.

There is reason to be skeptical that the current prescriptions for using VAMs can ever succeed in measuring teaching contributions well. Given that current models rely on highly constrained state tests that cannot accurately measure growth for large numbers of students, and they operate in a highly unequal educational and social system that cannot be manipulated away by statistical controls, these efforts may be doomed to eternal inaccuracy and bias. As Haertel (2013) explains:

> No statistical manipulation can assure fair comparisons of teachers working in very different schools, with very different students, under very different conditions. (p. 24)

Efforts to correct one set of problems often lead to others. For example, concerns about the large, uncontrolled differences among school settings have led to the use of school fixed effects in many models. This approach results in effectively ranking teachers against each other within a school. Thus, even if a school has worked hard to select, develop, and retain only effective teachers, some will, by requirement of the model, be labeled ineffective. The reverse is also true: In a school where all teachers are incompetent, some would be rated effective by emerging at the top of the within-school distribution. Furthermore, this strategy exacerbates Susan Moore Johnson's concern: Because the technology of VAM ranks teachers against each other relative to the gains they appear to produce for students, one teacher's gain is another's loss, thus creating disincentives for collaborative work.

Similarly, analysts have found that results can be skewed by incorrect links between teachers and students in data sets. However, Ballou and Springer (this issue) suggest that, if given the chance to make corrections, teachers may try to "cook" their rosters by dropping students unlikely to support higher value-added scores, especially if they can predict how biases in the models work. Trying to fix VAMs is rather like pushing on a balloon: The effort to correct one problem often creates another one that pops out somewhere else. Is it worth all of this trouble for results that may continue to be disappointing?

Most worrisome are studies suggesting that teachers' ratings are heavily influenced by the students they teach even after statistical models have tried to control for these influences. These range from falsification studies showing that teacher "effects" are as strong for classes teachers have not yet taught as they are for the ones they actually served (Rothstein, 2010) to studies showing that classroom characteristics are strong predictors of changes in value-added scores from year to year and class to class even after demographics and prior test scores are taken into account (Newton et al., 2010). The circumstance described below, in which two teachers' value-added ratings flip-flopped when they exchanged assignments, is not unusual:

> We had an 8th grade teacher, a very good teacher, the "real science guy" … [but] every year he showed low EVAAS growth. My principal flipped him with the 6th grade science teacher who was getting the highest EVAAS scores on campus. … [And] now the 6th grade teacher [is showing] no growth, but the 8th grade teacher who was sent down is getting the biggest bonuses on campus. (Amrein-Beardsley & Collins, 2012, p. 15)

The notion that there is a stable "teacher effect" that is a function of the teacher's teaching ability or effectiveness is called into question if the specific class or grade-level assignment is a stronger predictor of the value-added rating than the teacher.

Some argue that instability in VAM ratings could be overlooked if districts merely focus on those teachers who routinely score near the bottom of the distribution. However, such consistently low ratings may occur only because certain teachers consistently teach students whose achievement gains are not measured on the grade-level tests—new English learners, for example, or students in gifted and talented classes who have already maxed out on the tests. Horror stories about highly respected teachers whose students show large gains on other measures being denied tenure or threatened with dismissal because of low VAM ratings on the state test are occurring with greater frequency (cf., Amrein-Beardsley & Collins, 2012; Ferrette, 2013; Pallas, 2012).

The most tragic outcome will be if VAM measures are used to ensure a spread in the ratings of teachers so as to facilitate dismissals, but the teachers who are fired are not the "incompetent deadwood" imagined by advocates. Instead, they are the teachers working with the most challenging students in the most challenging contexts and those whose students are so far ahead of the curve the tests have no items to measure their gains, and perhaps those who eschew test prep in favor of more exciting, but less testable, learning experiences. If value-added measures continue to prove untrustworthy, the likelihood that they can be used to improve the quality of teaching, or of the teaching force, will be remote.

A Modest Proposal

What if, instead of insisting on the high-stakes use of a single approach to VAM as a significant percentage of teachers' ratings, policymakers were to acknowledge the limitations that have

been identified and allow educators to develop more thoughtful approaches to examining student learning in teacher evaluation? This might include sharing with practitioners honest information about imprecision and instability of the measures they receive, with instructions to use them cautiously, along with other evidence that can help paint a more complete picture of how students are learning in a teacher's classroom. An appropriate warning might alert educators to the fact that VAM ratings based on state tests are more likely to be informative for students already at grade level, and least likely to display the gains of students who are above or below grade level in their knowledge and skills. For these students, other measures will be needed.

What if teachers could create a collection of evidence about their students' learning that is appropriate for the curriculum and students being taught and targeted to goals the teacher is pursuing for improvement? In a given year, one teacher's evidence set might include gains on the vertically scaled Developmental Reading Assessment she administers to students, plus gains on the English language proficiency test for new English learners, and rubric scores on the beginning and end of the year essays her grade level team assigns and collectively scores.

Another teacher's evidence set might include the results of the AP test in Calculus with a pretest on key concepts in the course, plus pre- and posttests on a unit regarding the theory of limits which he aimed to improve this year, plus evidence from students' mathematics projects using trigonometry to estimate the distance of a major landmark from their home. VAM ratings from a state test might be included when appropriate, but they would not stand alone as though they offered incontrovertible evidence about teacher effectiveness.

Evaluation ratings would combine the evidence from multiple sources in a judgment model, as Massachusetts' plan does, using a matrix to combine and evaluate several pieces of student learning data, and then integrate that rating with those from observations and professional contributions. Teachers receive low or high ratings when multiple indicators point in the same direction. Rather than merely tallying up disparate percentages and urging administrators to align their observations with inscrutable VAM scores, this approach would identify teachers who warrant intervention while enabling pedagogical discussions among teachers and evaluators based on evidence that connects what teachers do with how their students learn. A number of studies suggest that teachers become more effective as they receive feedback from standards-based observations and as they develop ways to evaluate their students' learning in relation to their practice (Darling-Hammond, 2013).

If the objective is not just to rank teachers and slice off those at the bottom, irrespective of accuracy, but instead to support improvement while providing evidence needed for action, this modest proposal suggests we might make more headway by allowing educators to design systems that truly add value to their knowledge of how students are learning in relation to how teachers are teaching.

REFERENCES

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(12). Available at http://epaa.asu.edu/ojs/article/view/1096

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems [Special issue]. *Educational Researcher, 44*, 77–86.

Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/due-diligence

Corcoran, S. (2010). *Can teachers be evaluated by their students' test scores? Should they be?* Providence, RI: Annenberg Institute for School Reform, Brown University.

Darling-Hammond, L. (2013). *Getting teacher evaluation right*. New York: Teachers College Press.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. San Francisco: Jossey-Bass.

Ferrette, C. (2013, November 3). Great Neck teacher sues state over teacher evaluation system. *New York Newsday*. Available at http://www.newsday.com/long-island/education/great-neck-teacher-sues-state-over-teacher-evaluation-system-1.9581397

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce [Special issue]. *Educational Researcher, 44*, 87–95.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data [Special issue]. *Educational Researcher, 44*, 96–104.

Haertel, E. (2013). *Reliability and validity of inferences about teachers based on student test scores* (Angoff lecture). Princeton, NJ: Educational Testing Service.

Harris, D. N., & Anderson, A. (2011). *Bias of public sector worker performance monitoring: Theory and empirical evidence from middle school teachers*. Paper presented at the 2011 annual meeting of the Association for Education Finance and Policy.

Jackson, C. K. (2012, October). *Teacher quality at the high-school level: The importance of accounting for tracks* (National Bureau of Economic Research Working Paper No. 17722). Cambridge, MA: National Bureau of Economic Research.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students [Special issue]. *Educational Researcher, 44*, 105–116.

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school? [Special issue]. *Educational Researcher, 44*, 117–126.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23).

Pallas, A. (2012, May 15). The worst eighth-grade math teacher in New York City. *A Sociological Eye on Education*. Retrieved December 1, 2012 from http://eyeoned.org/content/the-worst-eighth-grade-math-teacher-in-new-york-city_326/

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1): 175–214.

Rothstein, J. (2011). *Review of "learning about teaching: Initial findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center.

Sass, T. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington DC: CALDER.

AUTHOR

LINDA DARLING-HAMMOND, EdD, is a Charles E. Ducommun Professor at Stanford University School of Education, 520 Galvez Mall, Stanford, CA 94035; *lindadh@stanford.edu*. Her research focuses on teaching quality, educational equity, and school reform.

Manuscript received January 26, 2015
Revisions received February 3, 2015, and February 7, 2015
Accepted February 9, 2015