

# *Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation*

JOHN P. PAPAY

*Brown University*

*In this article, John Papay argues that teacher evaluation tools should be assessed not only on their ability to measure teacher performance accurately, but also on how well they inform and support ongoing teacher development. He looks at two major approaches to teacher evaluation reform: value-added measures and standards-based evaluations. Papay analyzes these two approaches both as measurement tools and as professional development tools, illuminating the advantages, drawbacks, and untapped potential of each. In the process, attention is refocused towards a broader conception of the purpose of teacher evaluation.*

Over the past decade, consensus has been growing among teachers, administrators, and policy makers: teacher evaluation in the United States is broken and needs fixing. In school districts across the country, few teachers are evaluated regularly, and the evaluations that do occur are cursory. Not surprisingly, nearly all teachers succeed on these evaluations, and very few teachers are identified as unsatisfactory. These limitations have led to calls for reform, and districts across the country have struggled to identify and implement better evaluation systems.

In these conversations, however, there is little agreement about exactly what type of a system should replace the status quo. Recently, value-added models that purport to estimate a teacher's contribution to student test-score growth have grown in popularity, particularly among policy makers who like their explicit focus on student outcomes. Other districts have transformed the traditional system by introducing standards-based evaluations, rigorous and data-driven classroom observations in which expert evaluators assess a teacher's practice relative to explicit and well-defined district standards. Although both approaches present clear opportunities to improve teacher evaluation, each

has important limitations. Assessing their relative merits, and their ability to effect serious evaluation reform, requires a deeper understanding of how a performance evaluation system should function.

Evaluation systems can serve two main purposes. First, they can assess how effectively teachers are doing their jobs. In other words, they are measurement tools that districts can use to hold teachers accountable, removing teachers who do not meet the districts' standards and possibly rewarding top performers. Second, evaluations can provide valuable information to drive professional growth and, as such, can raise teacher effectiveness. As a formative professional development tool, evaluation provides feedback on teachers' instructional strengths and weaknesses, highlights areas for improvement, and supports teachers' continued development.

For the most part, policy debates regarding evaluation systems have revolved around the first of these purposes, focusing on evaluations as measurement instruments to assess teacher performance and hold teachers accountable. As measurement tools, teacher evaluations—both value-added models and standards-based observations—should be judged according to three criteria: are they unbiased, reliable, and valid? Policy makers have become enamored with value-added models because they are quantitative in nature and are seen as objective and inherently valid and reliable. Meanwhile, standards-based evaluations that rely on classroom observations are seen as subjective and bias-ridden. Research suggests that these two approaches are in fact more alike than different as measurement instruments, and both face serious concerns about bias, reliability, and validity.

Although developing an evaluation system that can assess teacher performance with strong validity is important, policy makers and researchers have focused on this purpose much too narrowly. If teacher evaluation is to improve student learning systematically, it must be used as a tool to promote continued teacher development. Using teacher evaluations in this manner holds much more promise for comprehensive change than identifying (and rewarding or sanctioning) the best and worst performers. Even if evaluation reform produces valid and reliable measures that policy makers use to alter the teaching force, relatively few teachers will be affected because few are identified as consistently high or low performing. By contrast, evaluation systems that improve instructional effectiveness can have a much broader impact. Research suggests that a rigorous evaluation program does boost teacher effectiveness and student achievement (Taylor & Tyler, 2011). In this regard, evaluation systems should be judged not only by their quality as measurement tools but also by the quality of the targeted feedback they provide and by their ability to drive continued instructional improvement.

In this article, I seek to reframe the debate about teacher evaluation reforms. I begin by examining the growing consensus for reform in teacher performance evaluation and describe these two approaches. I assess how well value-added estimates and standards-based observations work as measure-

ment instruments and professional development tools. I explain the concepts of bias, reliability, and validity, arguing that, as measurement tools, neither approach can provide unbiased, reliable, and valid assessments of teacher performance, but they both can represent clear improvements over the status quo. As professional development tools, I argue, standards-based protocols provide teachers with more meaningful feedback, but value-added data also provide some currently untapped opportunities to drive teacher growth. I then describe some key issues in implementation, particularly as districts seek to create systems that can serve both purposes. I conclude by arguing that developing better measurement tools is important, but it will not achieve the ultimate goals of transformational change. If we seek to produce widespread and systematic improvements in student learning, then efforts to reform teacher evaluation should refocus on continued teacher improvement. We should shift the debate from an argument about the best way to identify the top and bottom performers to a discussion of how best to use performance evaluation to improve teacher development and boost student learning in schools.

### Teacher Evaluation: A Growing Consensus for Reform

Policy makers, administrators, and teachers on all sides of the teacher evaluation debate acknowledge that teachers are critical to student learning and that not all teachers are equally effective. Quantitative research supports the now well-worn mantra that teachers are the most important school-level factor in promoting student achievement and shows that there is wide variation in teachers' abilities to raise student test scores (Aaronson, Barrow, & Sander, 2007; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). This variation is self-evident to educators: according to The New Teacher Project's *Widget Effect* report, 81 percent of all administrators and 57 percent of all teachers in the four districts studied reported that a tenured teacher in their school was not effective (Weisberg, Sexton, Mulhern, & Keeling, 2009).

These two points of consensus lead to a third: teacher evaluation in most districts is broken and must be reformed in order to improve instruction and teaching quality (Donaldson, 2009). Randi Weingarten (2010), president of the American Federation of Teachers, put it starkly: "Our system of evaluating teachers has never been adequate" (p. 3). It is no secret that few teachers are evaluated regularly. A recent report on the Boston Public Schools found that only half of all tenured teachers had been evaluated in the past two years (National Council on Teacher Quality, 2010). Furthermore, many of the evaluations that do occur consist only of so-called "drive-by" observations in which a principal stops into a classroom for a brief visit and indicates whether the teacher is "satisfactory" or "unsatisfactory" on a basic checklist of practices (Toch & Rothman, 2008).

The lack of effective evaluations leads to a “Lake Wobegon” problem where nearly all teachers are deemed “satisfactory” (Donaldson, 2009). For example, in a study of evaluation programs in twelve school districts, The New Teacher Project found that less than 1 percent of all teachers are rated as unsatisfactory (Weisberg et al., 2009). As a result, despite the recognition that some teachers may not be successful, almost no teachers are dismissed (Honawar, 2007; Tucker, 1997). Furthermore, very few teachers report that they get meaningful feedback—or any feedback at all (Weisberg et al., 2009). Thus, teacher evaluation must be improved. In assessing the possible alternatives, we must compare new approaches to the status quo: a system in which few teachers are evaluated, evaluations that do occur are brief and cursory, and nearly all teachers receive one piece of feedback—that their performance is “satisfactory.” In such a system, not only do administrators and policy makers gain no real information about teacher effectiveness, but teachers receive no meaningful feedback to help them improve their instructional practices.

There are strong sources of disagreement, though, about what strategies should be introduced to fix the problem. Recently, the technology available to evaluate teachers has improved dramatically. The development of comprehensive, longitudinal datasets and the rise of regular standardized testing have spurred the development of value-added models. These models can provide estimates of an individual teacher’s effectiveness in raising student test scores. At their heart, value-added models seek to predict how a student would have done on a test, using information such as their past test performance, other background characteristics, and characteristics of their peers and their school. The models then compare how a teacher’s students actually perform on the test to these predictions.

Value-added approaches are attractive to policy makers for several reasons. First, with the growing focus on test-based accountability, these measures directly assess student test-score growth. In other words, they explicitly focus on educational outputs rather than inputs. Because they are based on external assessments, they are seen as objective. And, with readily available datasets, these measures can be fairly easy and inexpensive to estimate. While value-added measures are not perfect, they represent an improvement over the current system. One of the value-added movement’s largest contributions is that it has focused attention on teacher effectiveness and raised serious questions about the status quo.

Standards-based evaluations build on the traditional model of teacher observations; however, this approach goes beyond simple classroom observations. In a rigorous system, the district develops a clear set of instructional standards and a detailed rubric that explains specific levels of performance for each standard. For example, Cincinnati, which implemented standards-based observations in 2000, has sixteen standards and thirty-two performance elements within those standards. Trained evaluators observe individual teachers

several times during the year, scripting lessons and matching evidence to the standards and the rubric. At the end of the cycle, these evaluators complete a summative assessment that provides detailed information about a comprehensive set of classroom practices. Evaluators rate teachers as “distinguished,” “proficient,” “basic,” and “unsatisfactory” (Kane, Taylor, Tyler, & Wooten, 2011). These evaluations rely on observations that include all teachers in the district, not just those in tested grades and subjects. The instructional standards explicitly examine teachers’ instructional interactions with students and students’ responses to their teachers (Johnson, Fiarman, Munger, Papay, & Qazilbash, 2009).

These standards-based evaluations have grown out of a burgeoning effort to define standards of instructional practice. Efforts like the National Board for Professional Teaching Standards and Charlotte Danielson’s Framework for Effective Teaching pushed this conversation forward, and districts have begun developing instructional standards based on these frameworks. A well-implemented standards-based evaluation system provides several advantages over traditional classroom observations. They afford a much richer view of a teacher’s instructional practice because evaluators visit classrooms several times over the course of the year. They are also based on clear evidence and standards, rather than administrators’ hunches or judgments. Evaluators must justify all assessments with the evidence that they have collected during the observation. Finally, unlike traditional observations, in which most teachers report getting little useful feedback (Weisberg et al., 2009), these evaluations provide rich information about instructional practices and how these practices meet the district’s standards.

### Assessing Two Evaluation Approaches

Understanding the advantages and limitations of different evaluation approaches requires us to examine these approaches in light of their potential purposes. Evaluations can be used as both measurement tools to assess performance and as professional development tools to improve instructional practice. Much of the policy discussion and research concerning teacher evaluation reform has focused on the first purpose: how we can assess teacher performance most effectively. But, the second is just as, if not more, important if evaluation reform is to produce systemic improvements in student learning. If we are to reframe our thinking to focus on professional development and systemic improvement, our assessment tools must be able to measure accurately and be unbiased, valid, and reliable. But they also must do much more than that. They must offer information that teachers can use to improve practice. In the following sections, I examine value-added models and standards-based observations as measurement tools and as professional development tools, reviewing the research on how they achieve these two goals.

### *Evaluation Approaches as Measurement Tools*

An evaluation system can help increase student learning by monitoring teacher performance and assessing how effective teachers are in the classroom. This is the traditional purview of employee evaluation and the main goal of most policy makers who argue for improved teacher evaluation systems. The theory of action holds that evaluations can improve instruction both by providing incentives for teachers to work hard (systemic improvement) and by removing the least effective teachers from the district (selection). Evaluation is thus a summative assessment that must provide a high-quality measure of how well teachers contribute to student learning.

In this regard, an evaluation is essentially a measurement tool. Advocates for different approaches regularly speak in the language of measurement to justify their position. For example, Battelle for Kids (2011), a leading provider of value-added modeling for states and school districts, claims that “while the statistical methodology used for value-added analysis is complex, the data produced are valid [and] reliable.” As a result, it is important to distinguish and understand three key concepts: bias, reliability, and validity (see Koretz, 2008, for a more detailed discussion).

A good measurement instrument is both unbiased (on average, it gives you the correct answer) and reliable (it gives you the same answer if you use it repeatedly). For example, consider a measurement tool that we encounter regularly: a thermometer. To be useful, a thermometer must not give us widely different readings each time we look at it (be reliable) and must not give us the wrong answer (be unbiased). Unlike reliability and bias, which are properties of the tool itself, validity is a property of the inference we hope to draw from that tool. An unbiased and reliable thermometer would allow us to make some valid inferences, perhaps indicating whether we need to wear a sweater. But the thermometer would not be valid to draw conclusions about whether we need to wear sunscreen. Thus, the thermometer is not “valid” or “invalid” in and of itself, but it may be valid or not for specific purposes. For a measurement tool to enable valid inferences, it must be both (relatively) reliable and free of bias.<sup>1</sup>

#### — Bias

Obviously, limiting bias in evaluation measures is an important concern; an evaluation tool is not useful if it does not give an accurate portrayal of a teacher’s performance. One typical concern about standards-based observations is that evaluators may not provide objective assessments of classroom practice because of underlying prejudices against the teacher. In a recent study, an administrator we spoke with voiced this concern, suggesting that principals may base judgments on outside information; she said that a teacher might argue that the results of an evaluation were unfair by saying, “Well, this is the principal who is just angry with me. She’s mad because I didn’t organize the Christmas party last year” (Papay & Johnson, in press).

Limiting bias in standards-based observations presents challenges because such observations rely on human judgments. Evaluators report that it is difficult to separate what they know of the teacher, or the teacher's contributions outside of the classroom, from their judgments of the teacher's instructional practice. However, having clear standards, using highly qualified and well-trained evaluators, and focusing on evidence can help remove much of the subjective bias. One evaluator we interviewed in Syracuse described the use of standards: "It's not just what I think is good teaching. It's not just my judgment. It is based on those performance indicators. That's what we're looking for" (Fiarman, Johnson, Munger, Papay, & Qazilbash, 2010, p. 14). Nonetheless, separating the personal from the professional can be difficult, and bias in these observation measures remains an important consideration.

By contrast, value-added models are based on objective test scores, not personal judgments. These test scores are typically from state standardized tests, which are graded by machine or by external scorers with no chance for systematic favoritism to certain teachers or schools. This can be an important advantage of value-added models; however, this alone does not eliminate bias, and the degree of bias may in fact be quite large. Possibly the largest threat that may bias value-added estimates is the extent to which value-added models can fully account for differences in student assignments. Value-added models typically account for a wide variety of student characteristics, explicitly comparing the performance of students with similar test-score histories. If value-added measures hope to isolate a teacher's contribution to student achievement growth, they must fully account for these differences in students taught, both within and across schools.<sup>2</sup>

The debate among researchers about the amount of bias that results from this nonrandom matching of students to classrooms has not been resolved. Kane and Staiger (2008) compare estimates of teacher effectiveness from a value-added model in Los Angeles to estimates derived from an experimental evaluation in which students were randomly assigned to teachers. Through random assignment, the researchers were able to account for differences among students, on average. They conclude that the results from the value-added measures approximate unbiased estimates. However, Rothstein (2010) argues that the sorting of students to teachers in North Carolina elementary schools produces considerable bias in value-added estimates. In particular, he finds that a student's fifth-grade teacher predicts their achievement gains in fourth grade (before the student has ever had the fifth-grade teacher), suggesting that some teachers tend to have certain types of students. The key challenge is that students are not randomly assigned to teachers, and the statistical controls used in value-added models may not sufficiently correct for this sorting.

Beyond student assignments, a second major source of bias comes from the ability of value-added models themselves to estimate a teacher's impact on stu-

dent performance. Value-added models rest on a set of substantive and technical assumptions that are typically not well-described, and these assumptions can substantially affect estimates of teacher effectiveness. For instance, attributing a student's mathematics test gains to her mathematics teacher may not fully recognize the contribution that a science teacher made to the student's mathematical knowledge. In many states, testing occurs in the middle of the spring semester rather than at the end of the year. Thus, any learning that happens (or does not happen) during the rest of the year or over the summer is attributed to the following year's teacher. Furthermore, the technical properties of the tests themselves, such as how they are scaled and whether they have ceiling effects, can make a difference (Ballou, 2009; Briggs & Weeks, 2009; Koedel & Betts, 2010). In tests with ceiling effects, high-performing students tend to earn a perfect score, and there is no way to differentiate between their levels of proficiency. As a result, value-added estimates for these students will not reflect their true improvement. These assumptions have been widely discussed in the value-added literature, but they have not been fully acknowledged by the policy community (for more detail, see McCaffrey et al., 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Reardon & Raudenbush, 2009; Todd & Wolpin, 2003). Thus, although advocates profess that value-added models isolate a teacher's contribution to student learning, these claims are likely too strong. Despite their quantitative nature, value-added models are not free of bias.

— Reliability

Reliability is a widely reported concern with classroom observations. There are two main issues. First, because high-quality observations are time consuming, evaluators must make judgments based on a relatively limited sample of instruction. A common criticism of traditional evaluation is that observations are announced, so teachers can prepare and execute an effective lesson on the day that they are observed. Second, because different evaluators may have different standards, achieving sufficient inter-rater reliability may be difficult. It is entirely possible that two different evaluators could rate the same teacher's practice differently. Teachers should not be rewarded (or punished) simply for having an easy (or tough) evaluator.

Although building reliability takes a substantial investment, it is possible. Incorporating multiple observations into an evaluation helps a great deal, particularly if the observations are unannounced. For example, in Cincinnati, Ohio, evaluators generally observe at least four complete lessons before making their final determinations. Furthermore, all evaluators must complete a comprehensive training to ensure that they interpret each standard similarly. Part of this training involves evaluating instruction on video-recorded lessons. Prospective evaluators need to achieve sufficient agreement on these lessons in order to be certified (Johnson et al., 2009; Kane et al., 2011). The district also invests heavily in ongoing training and professional development of evalu-

ators. Evaluators meet biweekly for additional training, where they continue to build a common understanding of the instructional standards and to calibrate ratings (Donaldson, 2009). Through such processes, reliability can improve.

Despite routine claims by policy makers who support value-added approaches that the measures are “valid and reliable,” there has been less empirical attention to the reliability of these estimates. What evidence exists suggests that there is cause for concern. Researchers have attempted to quantify the variability in value-added measures in several ways. First, they have examined whether teachers’ value-added estimates are similar from year to year. Although the specific results depend on the dataset and model used, most studies find moderate-sized year-to-year correlations that average between 0.4 and 0.5 (e.g., McCaffrey, Sass, Lockwood, & Mihaly, 2009). Because a teacher’s performance likely varies from year to year, it is not clear exactly how to interpret these findings. However, these figures would represent substantial changes in teacher effectiveness from one year to the next.

These results have led researchers, even those who advocate the use of value-added approaches, to call for using multiple years’ worth of data to construct an estimate for an individual teacher. For example, in a report issued by the Brookings Institution, a group of experts who support the use of value-added measures argue that “any practical application of value-added measures . . . should include multiple years of value-added data in combination with other sources of information to increase reliability and validity” (Glazer et al., 2010, p. 6). Such best practices are important to follow, but they severely restrict the utility of value-added approaches for annual evaluation.

Researchers have also examined value-added estimates from two different tests of the same content area in the same year. The idea is that if the tests measure similar material, estimates of a teacher’s effectiveness using either test should be quite similar. However, even when they include multiple years of data, these correlations typically range between 0.3 and 0.5 (Corcoran, Jennings, & Beveridge, 2011; Gates Foundation, 2010; Lockwood et al., 2007; Papay, 2011). These estimates are not sufficiently reliable to classify the highest- and lowest-performing teachers consistently. For instance, in one study, 30 percent of the teachers who are rated as below average using one student achievement test would have scored in the top 25 percent of teachers using the other test (Papay, 2011).

#### — Validity

Validity is the most important of these three criteria, but it is also the most difficult to assess. Again, validity is a property of a conclusion that we hope to draw from a measure, not a property of the measure itself. In most cases, validity requires the measure to be unbiased and reliable—if the measure tends to give the wrong answer or is unreliable, it is not of much use for any inference. One key challenge in assessing the validity of evaluation measures comes in defining what districts hope to measure. Evaluating whether teachers promote

*student learning* is one thing, while evaluating whether they raise *student test scores* is another.

Although policy makers often conflate these two goals, they are quite different, and this distinction has important implications for assessing both standards-based observations and value-added approaches. Value-added approaches are popular in part because they focus directly on educational outcomes rather than process—this is a key advantage of these measures. Obviously, though, value-added models only examine student test scores. If student test-score growth itself does not reflect actual learning, then these estimates will not be valid for drawing inferences about teacher performance. In other words, value-added estimates are only as good as the tests on which they are based. And, clearly, they do not reflect other learning that is not captured on standardized tests. Currently, value-added estimates are only practical for teachers of annually tested subjects (typically mathematics and English language arts in grades 4–8). High school, early elementary school, history, science, and arts teachers are thus excluded.

Furthermore, actions by teachers can threaten the validity of the measurement. Teachers may be able to raise student test scores in a variety of ways that do not promote student learning, such as teaching to the test or cheating (Figlio & Getzler, 2006; Jacob & Levitt, 2003; Koretz, 2008). Recent scandals in Washington, Atlanta, Philadelphia, and other cities have found evidence of test cheating by teachers and administrators. More commonly, administrators and teachers may narrow the curriculum to specific types of questions that frequently appear on state tests. In both cases, student test scores may rise, but the test evidence may not support inferences about how much students have been learning. This is a common problem when districts attach incentives to evaluations that do not measure what they hope to improve; in the descriptive title to his classic paper, Kerr (1975) calls this the “folly of rewarding A, while hoping for B.”

Standards-based observations face similar challenges in seeking to evaluate a teacher’s contributions to student learning. Fenstermacher and Richardson (2005) draw a useful distinction between good teaching and successful teaching. Good teaching involves using practices that are developmentally appropriate and pedagogically sound, while successful teaching produces results.<sup>3</sup> Standards-based observations must distinguish between teachers who simply use good practices and those who use these practices effectively to promote student learning. In other words, an effective standards-based observation system will evaluate not simply the teacher but also teacher-student interactions.<sup>4</sup>

In recent years, several researchers have attempted to “validate” observational measures by comparing teachers’ evaluation ratings to value-added estimates. Clearly, this approach may not necessarily validate the standards-based observations because the inferences from value-added models themselves may not be correct. However, the exercise proves useful because, to the extent that standards-based observations reveal practices that are effective at improving

student knowledge, they should be related to increased test scores. Indeed, standards-based evaluations are relatively strong predictors of teachers' value-added measures (Grossman et al., 2010; Hill, Kapitula, & Umland, 2011). In fact, using data from Cincinnati, Kane and colleagues (2011) found that a teacher's standards-based evaluation rating actually predicts student test performance above and beyond that teacher's value-added rating.

Interestingly, the correlations between standards-based evaluation ratings and teacher value-added estimates are quite similar to those between teacher value-added estimates that use different student tests in the same subject. Thus, while supporters argue for value-added models because they measure student outcomes directly, this rhetoric is not fully supported by the research evidence. In fact, the evidence suggests that standards-based observations may measure student learning—as captured in test-score growth—just as well as value-added measures do.

#### *Evaluation Approaches as Professional Development Tools*

Most attention from policy makers and the research community has centered on assessing and improving teacher evaluations as measurement tools. Although this goal is important, it ignores a key purpose to teacher evaluation: to improve instruction by developing teachers' instructional capacity and effectiveness. Such a system can raise aggregate performance through systemic improvement rather than selection of teachers out of the profession. By identifying areas in which a teacher succeeds or fails, an evaluation enables teachers to leverage areas of strength and remediate areas of weakness. Rich and specific feedback can promote such improvement. Here, the evaluation is essentially a formative assessment. The evaluation system can and should be seen as a professional development tool and should be evaluated on its ability to raise instructional proficiency and student learning.

Assessing the prospects of an evaluation system as a tool for continuous instructional improvement requires not only examining its reliability, validity, and bias but also identifying the system's prospects for driving instructional change. A common criticism of value-added approaches is that simply receiving an evaluation score does not tell teachers how to improve. As implemented in most districts, value-added measures do have this serious limitation. But, with the rich performance data embedded in district datasets, more detailed analysis is possible. Some districts already use disaggregated student data to help target instruction. Using measures of student growth could help refine this process by providing teachers with a more accurate idea of the areas in which they are having success (or difficulty) with their students. For example, districts could provide data about teacher effectiveness on certain types of test questions or with certain types of students (e.g., some teachers may teach better to boys than girls or to high-performing students than lower-performing students). This information could prove useful as teachers attempt to refine their practice.

Furthermore, viewing teacher evaluation as a measurement tool necessarily places the focus on the individual. But, schools are collections of teachers who work together for a common purpose. As a professional development tool, evaluation can prove useful in helping build organizational capacity. For example, principals can use evaluation data to identify broader areas of instructional strength and weakness in the school. They can then target resources appropriately and leverage existing teachers who have had success in certain areas to share their knowledge. Using value-added data in this more systematic manner can help build organizational capacity. Unfortunately, because few districts have begun to use targeted value-added results to improve instruction directly, no research exists about the effect of such practices.

Although value-added measures have unrealized potential as professional development tools, standards-based observations provide more direct and specific feedback about a teacher's instructional practice. The detailed observation reports and summative evaluations include meaningful information about where teachers are succeeding and where they can improve. The performance standards and rubrics also offer teachers a clear sense of what practices they need to adopt in order to succeed on the evaluation. Thus, these standards-based evaluations can provide teachers with a clear "line of sight" between their current practices and what they need to do to improve (Lawler, 1990). As described above, administrators can also use information from standards-based evaluations to examine the organizational strengths and weaknesses in the school.

There is mounting evidence that rigorous, standards-based evaluations can improve teacher effectiveness. Strikingly, Taylor and Tyler (2011) find that midcareer teachers who complete Cincinnati's standards-based evaluation system improve their performance not only during the year in which they are evaluated, but in subsequent years as well. These improvements are quite substantial and suggest that teachers are learning during the evaluation and changing their ongoing instructional practices as a result.<sup>5</sup> Regardless of whether districts employ value-added or standards-based evaluation, teacher instructional capacity is strengthened through formative evaluation processes.

### Implementing an Evaluation Program

Whether districts use value-added models, standards-based observations, a combination, or some other approach entirely, thoughtful design and careful implementation are critical to the evaluation system's success. Both value-added approaches and standards-based evaluations require a commitment to invest resources in teacher evaluation, although standards-based observations are much more resource intensive. A thorough discussion of these issues goes beyond the scope of this article, but a few comments are in order about both types of systems.

*Standards-Based Observations*

A high-quality, standards-based evaluation system requires rigorous instructional standards with clear rubrics that define success on these standards. There are several models of well-crafted evaluation standards and rubrics, but these are not one-size-fits-all approaches that districts can simply adopt. Instead, each district must adapt these existing models to their local context and work carefully with both administrators and teachers to develop understanding, buy-in, and trust. There should also be standards of practice for the evaluation itself, including clear expectations about the level of evidence required to make a summative assessment and the extent of feedback provided to teachers. For example, a rigorous evaluation should incorporate data from multiple observations throughout the year to ensure sufficient reliability.

Furthermore, evaluators must be well trained, knowledgeable about effective teaching practices as defined in the standards, and able to analyze observed practices to determine how well teachers are meeting these standards. The importance of developing high-quality evaluators and the challenges they will face must not be underestimated. In a recent study of an evaluation program, we found that a key limiting factor was principals' unwillingness to identify teachers as not meeting standards; telling teachers they are not doing a good job is not easy work (Johnson et al., 2009). Effective evaluators must be willing to provide tough assessments and to make judgments about the practice, not the person. They must also be expert in providing rich, meaningful, and actionable feedback to the teachers they evaluate. Although the traditional model relies on administrators to conduct evaluation, several districts across the country have experimented quite successfully with identifying expert peers to serve as evaluators (Johnson et al., 2009). Regardless of who serves in the role, all evaluators must be trained and supported.

Not surprisingly, implementing such a system cannot be done cheaply. Planning and development take time and money, and operating a rigorous standards-based evaluation system is quite resource intensive. Many of the flaws in the status quo system arise because teachers are never evaluated. Simply adopting rigorous standards and protocols will not change this practice. Instead, evaluators need time to do this work well. In districts that use administrators as evaluators, that means providing additional support to relieve daily administrative tasks. In many schools, principals simply do not have the time to devote to conducting evaluations. Using expert peer teachers helps distribute the burden, although providing sufficient time is still expensive. That said, several districts—such as Cincinnati, Ohio, Montgomery County, Maryland, and Washington, D.C.—have implemented standards-based evaluation programs. In part, these districts use creative practices to limit some of the program's costs. In Cincinnati and Montgomery County, for example, veteran teachers who have had successful evaluations are not reevaluated every year.

Rather than doing a cursory annual evaluation, as is typical in many schools, these districts invest in rigorous evaluations every three to five years.

### *Value-Added Models*

Implementation also remains a key challenge for value-added evaluations, even though the estimates are quantitative in nature and are often calculated by a third-party analyst or firm. As described above, value-added estimates are sensitive to a variety of assumptions that analysts must make. Which assumptions a district chooses to make can determine how effective an individual teacher appears to be. As a result, policy makers, administrators, and teachers should spend sufficient time to understand the key assumptions behind these measures and should make informed decisions about these important analytical choices. Local officials should make these decisions with an understanding not only of how they will affect teacher ratings but also of how they will change the incentives facing teachers. For example, some value-added models statistically control for student-level demographic characteristics, while others do not. This decision may change the types of students that teachers prefer to teach.

There are also several practical limitations that restrict the broad-based application of value-added models. Most obviously, with standardized testing practices in place in most states, fewer than one in three teachers works in a grade or subject area that supports value-added analysis (typically English language arts and mathematics teachers in grades 4–8). Second, these estimates are only as good as the tests on which they are based. Because many state tests are designed to measure whether students are proficient, they may not be particularly good for evaluating teachers. Finally, because value-added estimates rest on the test data, principals need to wait at least until test results are available during the summer before assessing performance. And, the need to include multiple years of data to have sufficiently reliable estimates limits their use for annual teacher evaluation.

Although calculating value-added estimates themselves is not expensive, the infrastructure they rely on can be costly. For example, districts must work carefully to ensure that their data collection systems are up-to-date and accurate. For a data system to be used for such high-stakes purposes as teacher evaluation, it must meet levels of accuracy not currently found in most districts. If teachers' estimates are based on students who are not in their classes or on inaccurate student test scores, the evaluation system will not be effective.

If either system is to be used effectively as a professional development tool, the district must create strong structures to support teachers in using their evaluation results to improve. In a standards-based evaluation system, the evaluator may be able to provide this support. However, with a value-added system, the district must invest in personnel who can help teachers not only make sense of their ratings but use these data to inform their instructional practice.

## Conclusion: Refocusing the Teacher Evaluation Debate

The current evaluation system in place in many districts is clearly ineffective. Evaluations that do not happen or that consist of “drive-by” observations with no constructive feedback serve little purpose. Teachers see these as rituals that must be endured, not as professional evaluations designed to ensure quality and boost performance. If implemented well—certainly a big *if*—both value-added models and standards-based observations can surpass current evaluation practices in place in many districts as measurement instruments and as professional development tools. As measurement tools, both approaches face challenges in design, implementation, and interpretation; however, they both surpass an existing system in which teachers either are not assessed or are all rated as satisfactory. As professional development tools, standards-based evaluations have more promise, but even value-added models can be an improvement over a system in which teachers receive no meaningful feedback. However, in both cases, the relative effectiveness of a new system depends critically on the quality of implementation.

In these debates about evaluation reform, policy makers seem to be taking as their main priority the development of a better system to measure teacher performance. This is the traditional purview of evaluation, and it is clearly an important baseline. In any district, some teachers likely do not belong in the classroom, and administrators need to have a system in place that can hold those teachers accountable. A rigorous evaluation system should serve as a basis for dismissal if teachers cannot meet the district’s standards after receiving sufficient support. But in considering how to use evaluations to play this role, policy makers must recognize the important challenges around bias, reliability, and validity in both value-added models and standards-based observations.

Some of these challenges can be mitigated. For example, using multiple years of data in value-added models and investing heavily in evaluator training to boost standardization can both increase reliability. Furthermore, recognition of these limitations has led many analysts to call for using test-score data as only one of multiple measures of teacher performance. This approach is seen in places like Washington, D.C., whose IMPACT evaluation system gives each teacher a rating based on the following components: 50 percent value-added, 35 percent standards-based observation, 10 percent commitment to school community, and 5 percent school value-added.<sup>6</sup> Value-added measures and standards-based evaluation each have a role in evaluating teachers, and both can provide useful and important data. How to combine these different components in order to provide the best assessment of teacher effectiveness is not well understood and is an important area for future research.<sup>7</sup>

However, this entire framework—both in policy making and in research—continues to focus on evaluation simply as a measurement tool designed to identify the most and least effective teachers in a school. An effective evalua-

tion system must be much more than an assessment of teachers. By focusing on evaluation as a measurement tool to hold teachers accountable, the policy discussion has lagged well behind what is necessary to really “move the needle” on student instruction. This is particularly true given the scope of this challenge. With nearly three million teachers in the United States, rapid improvements in instructional effectiveness will not be possible by simply replacing low-performing teachers. Removing a few underperforming teachers from the classroom can certainly help, but it does not go nearly far enough. Instead, for evaluation to realize its potential as widespread instructional reform, it must work to raise the performance of all teachers in the system.

Evaluations must provide teachers with a clear understanding not only of their current success or failure, but also of the practices they need to develop to become more successful with their students. There are few examples of current systems that effectively combine these two purposes, effectively measuring teacher performance and providing feedback to help them improve. Peer Assistance and Review (PAR) programs, in place in dozens of districts across the country, have shown that this dual purpose is possible (Johnson et al., 2009). Several districts with PAR, including Cincinnati and Montgomery County, have applied these approaches for all teachers, developing rigorous evaluation systems that assess teachers’ current performance, hold them accountable for that performance, and shed light on practices that can be improved.

Research suggests that teachers can and do improve with specific and meaningful feedback (Taylor & Tyler, 2011). This type of professional learning underpins the best employee evaluations in many private-sector jobs. The current attention, among both education policy makers and scholars, to developing evaluation systems that serve as better measurement tools, then, is somewhat misguided. Although assessment is clearly an important goal, the ability of a system to promote continued teacher development should be a much greater priority. In the coming years, debates about the development of teacher evaluation systems will likely continue to take a prominent place in education reform discussions. These conversations will have the greatest impact on students and student learning if they focus on evaluation as a professional development tool to raise the instructional quality of existing teachers.

## Notes

1. Harris (2008) argues that the degree of validity (and hence the degree of bias and reliability) required to make a measure useful depends on the purpose for which you plan to use the tool and the tool’s costs. For low-stakes uses, we might be willing to accept measures that afforded less valid inferences about our question of interest than we would for high-stakes uses.
2. For more on value-added measures as causal estimates of teacher effectiveness, see Reardon and Raudenbush (2009) and Rubin, Stuart, and Zanutto (2004).
3. *Results* refers to a measurable outcome of interest to the district, such as student test scores. Fenstermacher and Richardson (2005) note that successful teaching alone is

not sufficient—many practices can produce results but not be aligned with good practice (one example they use is corporal punishment). Instead, effective teaching is both good and successful.

4. The Cincinnati TES provides an example of such a system. The performance standards not only include elements addressing how teachers prepare for class, their content knowledge, and their instructional strategies, but they also assess how teachers engage with their students. For instance, standard 2.1A explicitly focuses on “teacher interaction with students.” Although such practices get standards-based observations closer to the goal of measuring student learning, these evaluations do not focus directly on educational outputs.
5. That teachers improve not only during their evaluation year but also in subsequent years is an important finding. Standard principal-agent theory of supervision holds that employees will work hard only when they are monitored and thus have an incentive to do so. This line of reasoning would suggest that teachers should only improve when they are being evaluated carefully and that stronger incentives would improve teacher performance even more. However, a competing explanation holds that teachers are motivated largely by intrinsic goals, such as seeing their students succeed (Lortie, 1975). Here, student performance may not necessarily increase as a result of stronger incentives for teachers, because the teachers may not have the knowledge or capacity to improve (Elmore, 2003). In other words, they may not know what to do differently. Participating in an evaluation process that provides clear performance feedback thus could improve teacher performance not only when they are being monitored but also in subsequent years as teachers continue to apply what they have learned.
6. For teachers in nontested grades and subjects, the remaining components take on greater weight. For more information, see <http://www.dc.gov/DCPS/impact>.
7. See Glazerman et al. (2010) and Donaldson (2009) for an overview. For a more technical discussion, see Douglas (2007) and a 2003 special issue of *Educational Measurement* (Benson, 2003).

## References

- Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1), 95–135.
- Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, 4(4), 351–383.
- Battelle for Kids. (2011). *Value-added FAQs*. Retrieved from [http://portal.battelleforkids.org/tennessee/Resources/value\\_added\\_faq.html](http://portal.battelleforkids.org/tennessee/Resources/value_added_faq.html)
- Benson, J. (Ed.). (2003). Multiple perspectives on multiple measures. [Special issue]. *Educational Measurement: Issues and Practice*, 22(2).
- Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical scale score. *Education Finance and Policy*, 4(4), 384–414.
- Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). Teacher effectiveness on high- and low-stakes tests. Retrieved from [https://files.nyu.edu/sc129/public/papers/corcoran\\_jennings\\_beveridge\\_2011\\_wkg\\_teacher\\_effects.pdf](https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf)
- Donaldson, M. (2009). *So long Lake Wobegon? Using teacher evaluation to improve teacher quality*. Washington, DC: Center for American Progress. Retrieved from [http://www.americanprogress.org/issues/2009/06/pdf/teacher\\_evaluation.pdf](http://www.americanprogress.org/issues/2009/06/pdf/teacher_evaluation.pdf)
- Douglas, K. M. (2007). *A general method for estimating the classification reliability of complex decisions based on configurational combinations of multiple assessment scores*. Unpublished doctoral dissertation, University of Maryland, College Park.
- Elmore, R. F. (2003). *School reform from the inside out*. Cambridge, MA: Harvard Education Press.

- Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record*, 107(1), 186–213.
- Fiarman, S. E., Johnson, S. M., Munger, M. S., Papay, J. P., & Qazilbash, E. K. (2010). *Teachers leading teachers: The experiences of Peer Assistance and Review Consulting Teachers*. Working Paper, Project on the Next Generation of Teachers, Harvard Graduate School of Education. Retrieved from [http://www.gse.harvard.edu/~ngt/par/resources/SEF\\_AERA\\_2009.pdf](http://www.gse.harvard.edu/~ngt/par/resources/SEF_AERA_2009.pdf)
- Figlio, D. N., & Getzler, L. (2006). Accountability, ability and disability: Gaming the system? In T. Gronberg & D. Jansen (Eds.), *Improving school accountability—check-ups or choice?* (*Advances in microeconomics, vol. 14*) (pp. 35–49). Amsterdam: Elsevier.
- Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. MET Project Research Paper, Bill & Melinda Gates Foundation. Retrieved from [http://www.metproject.org/downloads/Preliminary\\_Findings-Research\\_Paper.pdf](http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf)
- Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Brown Center on Education Policy, Brookings Institution. Retrieved from [http://www.brookings.edu/~media/Files/rc/reports/2010/1117\\_evaluating\\_teachers/1117\\_evaluating\\_teachers.pdf](http://www.brookings.edu/~media/Files/rc/reports/2010/1117_evaluating_teachers/1117_evaluating_teachers.pdf)
- Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. Working Paper 45, National Center for Analysis of Longitudinal Data in Education Research. Retrieved from [http://www.caldercenter.org/upload/CALDERWorkPaper\\_45.pdf](http://www.caldercenter.org/upload/CALDERWorkPaper_45.pdf)
- Harris, D. (2008). The policy uses and policy validity of value-added and other teacher quality measures. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teacher quality* (pp. 99–130). Thousand Oaks, CA: Sage.
- Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794–831.
- Honawar, V. (2007). New York City taps lawyers to weed out bad teachers. *Education Week*, 27(14), 13.
- Jacob, B. A., & Levitt, S. D. (2003). Catching cheating teachers: The results of an unusual experiment in implementing theory. In W. G. Gale & J. R. Pack (Eds.), *Brookings-Wharton papers on urban affairs* (pp. 185–209). Washington, DC: Brookings Institution Press.
- Johnson, S. M., Fiarman, S. E., Munger, M. S., Papay, J. P., & Qazilbash, E. K. (2009). *A user's guide to Peer Assistance and Review*. Retrieved from <http://www.gse.harvard.edu/~ngt/par>
- Kane, T. J., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper 14607). Cambridge, MA: National Bureau of Economic Research.
- Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practice using student achievement data. *Journal of Human Resources*, 46(3), 587–613.
- Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4), 769–783.
- Koedel, C., & Betts, J. R. (2010). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, 5(1), 54–81.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.
- Lawler, E. E. (1990). *Strategic pay: Aligning organizational strategies and pay systems*. San Francisco: Jossey-Bass.

- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.
- Lortie, D. C. (1975). *Schoolteacher*. Chicago: University of Chicago Press.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.
- McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.
- National Council on Teacher Quality. (2010). *Human capital in Boston Public Schools: Rethinking how to attract, develop, and retain effective teachers*. Retrieved from [http://www.nctq.org/p/publications/docs/nctq\\_boston\\_human\\_capital.pdf](http://www.nctq.org/p/publications/docs/nctq_boston_human_capital.pdf)
- Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193.
- Papay, J. P., & Johnson, S. M. (in press). Is PAR a good investment? Understanding the costs and benefits of teacher Peer Assistance and Review programs. *Education Policy*.
- Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492–519.
- Rivkin, S., Hanushek, R., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.
- Rockoff, J. E. (2004). The impact of individual teachers of student achievement: Evidence from panel data. *American Economic Review, Papers and Proceedings, 94*(2), 247–252.
- Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175–214.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.
- Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (NBER Working Paper 16877). Cambridge, MA: National Bureau of Economic Research.
- Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.
- Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*(485), F3–33.
- Tucker, P. (1997). Lake Wobegon: Where all teachers are competent (or have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*(2), 103–126.
- Weingarten, R. (2010). *A new path forward: Four approaches to quality teaching and better schools*. Retrieved from [http://aft.3cdn.net/227d12e668432ca48e\\_twm6b90k1.pdf](http://aft.3cdn.net/227d12e668432ca48e_twm6b90k1.pdf)
- Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from <http://widgeteffect.org/>

This article has been reprinted with permission of the *Harvard Educational Review* (ISSN 0017-8055) for personal use only. Posting on a public website or on a listserv is not allowed. Any other use, print or electronic, will require written permission from the *Review*. You may subscribe to *HER* at [www.harvardeducationalreview.org](http://www.harvardeducationalreview.org). *HER* is published quarterly by the Harvard Education Publishing Group, 8 Story Street, Cambridge, MA 02138, tel. 617-495-3432. Copyright © by the President and Fellows of Harvard College. All rights reserved.