



Prioritizing Teaching Quality in a New System of Teacher Evaluation

By Heather Hill and Corinne Herlihy

Teachers are the most important school-level factor in student success—but as any parent knows, all teachers are not created equal. Reforms to the current quite cursory teacher evaluation system, if done well, have the potential to remove the worst-performing teachers and, even more important, to assist the majority in improving their craft. However, the US educational system often cannibalizes its own innovations, destroying their potential with a steady drip of rules, regulations, bureaucracy, and accommodations to the status quo. Because that status quo sets an unacceptably low bar for teaching quality, missing this opportunity now means new generations of students may suffer mediocre—or worse—classrooms.

Where We Are

Political will to overturn the existing teacher evaluation system has been growing for almost two decades, culminating recently in the federal Race to the Top legislation and district initiatives such as Washington, DC's IMPACT program, under which more than 280 teachers have lost jobs because of poor evaluations.¹ Initial skirmishes began in the late 1990s, when advocates of value-added statistical models that produce estimates of the student growth attributable to specific teachers touted their potential use in personnel decisions, including merit pay and termination.² These models soon fell under harsh critique, however, because teacher scores are often inconsistent from year to year, even though most believe teaching quality to be a relatively stable individual trait. Scores are vulnerable to bias stemming from student assignment to teachers and may be affected by student

access to after-school tutoring, parental help, and spillover effects from instruction in other subjects and prior years.³

Key points in this Outlook:

- Because classroom teaching is the most direct influence on student learning, we must prioritize the quality of teaching—not “teacher quality”—and insist on metrics more meaningful than the current “culture of nice” that gives 97 percent of teachers a rating of satisfactory or above.
- To address objections and shortcomings related to recent reform efforts, states and districts need to design a system of teacher evaluation that works with existing policies to improve teaching and learning.
- Reforming the evaluation system will bring about the greatest success not through carrots and sticks but through resources to help teachers improve their craft.

Heather Hill (heather_hill@gse.harvard.edu) is an associate professor at the Harvard Graduate School of Education with a focus on measuring mathematics instruction. Corinne Herlihy (corinne_herlihy@gse.harvard.edu) is the project director for the National Center for Teacher Effectiveness in the Center for Education Policy Research at Harvard University.

Despite these shortcomings, advocates of value-added-based teacher evaluation argue it is an improvement over the status quo, and evidence suggests they are correct. Most districts' evaluation systems suffer from a Lake Wobegon effect in which all teachers are above average. Principals' classroom evaluations are often cursory and, according to many, highly subjective. Until recently, objective criteria such as student gains were not included at all. And in many districts, identifying and terminating poor teachers has been difficult under union rules. Recent popular and scholarly attention to these realities has increased the current pressure on districts to develop more accurate and sophisticated teacher evaluation systems.

We argue, however, that any endeavor to overturn the status-quo teacher evaluation system will have to contend with another serious problem: American shortsightedness regarding teaching quality. For far too long, definitions of high-quality teaching have been local, variable, and superficial—and often focused heavily on matters other than instruction itself. District evaluation criteria often list dozens of elements, yet only a handful of those elements cover classroom work with students. And many of those lean heavily on superficial aspects of teaching quality, such as whether the teacher recorded the lesson objective on the board or asked three open-ended questions.

This phenomenon is not limited to teacher evaluation systems. Recent news stories about teachers imperiled by budget cuts, for instance, highlight teachers' attendance records, grading practices, lesson-planning records, and helpfulness to students. Often, these stories say little or nothing about the technical skills and expertise that render specific teachers effective or ineffective with students. This—as well as other problematic aspects of the system—must change for reform efforts to be effective.⁴

Where Do We Go from Here?

To effect such change, we argue that states and districts need to design a system of teacher evaluation that works in concert with existing policies aimed at improving teaching and learning. Several key elements characterize a successful teacher evaluation system:

- The system will provide teacher scores that accurately represent their skill and capacity in teaching and helping students learn;

- These scores will support key decisions, including termination but more important, tenure and the design of professional growth opportunities;
- The system will work efficiently, concentrating resources where they can make a difference;
- The system will provide useful feedback to all stakeholders—principals, instructional coaches, district staff, and most important, teachers themselves.

We highlight some design principles for such a system in the paragraphs that follow.

Prioritize the quality of teaching—not “teacher quality.”

Our first principle flows logically from the diagnosis above: policymakers should invest in a system that judges individuals directly on *teaching*, not *teacher*, quality. Teaching quality can be defined as the complex set of knowledge, skills, and judgments that comprise teachers' everyday work, such as conveying content accurately yet also at a level that can be understood by students, implementing cognitively challenging lessons and tasks, and diagnosing and remediating student misunderstandings.

Any endeavor to overturn the status-quo teacher evaluation system will have to contend with a serious problem: American shortsightedness regarding teaching quality.

There are several reasons behind this recommendation. First, classroom teaching is the most immediate influence on student learning—what and how teachers teach cannot help but heavily influence students' skills and abilities. Second, existing observational research suggests wide and sometimes alarming variability in teaching quality.⁵ Third, focusing evaluation systems on teaching itself creates incentives for teachers to improve a factor they directly control, unlike test-score outcomes that may be mediated by external factors such as student assignment to classrooms, assistance from parents, or extracurricular tutoring. This focus on teaching quality may help state policymakers encourage instructional improvement and even mitigate against the kind of gaming or cheating on high-stakes tests that has been recently uncovered in Atlanta and Pennsylvania.

For state or district officials, this means deepening the commitment to understanding and promoting good teaching. Many districts started down this road in the 1990s but quickly detoured toward worrying about how to improve student outcomes on standardized assessments. Moving to better develop the vision of good teaching and using this as a yardstick is a subtle yet important part of the work policymakers must now engage in.

Use multiple measures. Although we believe teaching quality should play a major role in new teacher evaluation metrics, we would be mistaken to say it should be the only factor. In fact, policymakers are already designing systems in which teachers' scores reflect their performance on multiple criteria beyond teaching, including producing gains in student achievement, contributing to the school community, and advocating for students and families. Such metrics produce teacher scores that average multiple elements, providing a balanced set of incentives for teachers.

However, we caution that variation must exist in the range of scores associated with each measure in the metric. Consider a district, for example, that uses value-added scores but has retained its old, everyone-is-above-average teacher observation system. Because there is so little variation in observation outcomes, teachers' overall evaluation will depend heavily—probably, too heavily—on the value-added scores. This again suggests the need for a meaningful teacher observation instrument that captures the substantial variation that exists in teaching quality.

Use value-added scores wisely. Although evidence suggests scores from value-added models are not sufficiently reliable and unbiased to use alone in high-stakes decisions, they do carry objective information that districts and schools can use to great advantage. Teachers with low value-added scores, for instance, should be flagged for extra scrutiny under the classroom observation component of the evaluation system. Teachers with high value-added scores should be observed and then, if appropriate, encouraged to act as mentors and peer coaches. In this way, this relatively inexpensive source of information—these models use only existing district and state data and are not labor-intensive—can provide a good starting point for many of the personnel decisions facing school leaders.

Improve existing observation instruments. As we have noted, many tools used to observe and evaluate teachers either focus little on teaching or measure its more

superficial aspects, such as the presence of a written lesson objective or three higher-order questions in instruction. Such observation instruments are easily gamed and likely not indicative of quality. Policymakers might instead turn to next-generation instruments capable of handling higher-inference judgments about teaching quality. For instance, some instruments ask observers to rate whether the lesson objective is clearly developed over the course of the tasks and activities students complete or whether once an open-ended question has been asked, the teacher can respond productively to students' answers.

Developing or choosing a high-quality observation instrument is only the first step; designing data-collection procedures for that instrument is the real work states face over the next several years.

Several instruments that measure such complex competencies have been developed, were validated in small-scale studies, and are now being tested on a large scale.⁶ Despite the existence of these well-vetted observation instruments, an informal poll of nine states that are moving forward with new evaluation systems reveals that most are either developing observation instruments from scratch or customizing one or two widely used existing instruments.⁷

State policymakers suggest that locally generated instruments may increase buy-in on the part of teachers and reflect local norms and contexts. We argue, however, that the costs of such a decision outweigh the benefits. Locally developed state standards for student learning have been widely vilified because of the resulting inequities for students; similarly, little reason exists to believe that good teaching looks tremendously different in Mississippi than in Massachusetts—or that bad teaching should be excused based on geography. But more important are the logistics involved: moving from a general vision of good teaching to sharp, precisely worded written indicators is time-consuming, involving months, if not years, of labor. Evaluation systems will fare better, with both teachers and the courts, if they provide evidence of score reliability and validity. Developing such evidence is costly and technically challenging—another reason to rely on existing, well-established instruments.

Worry less about the observation instrument and more about the observation system. Developing or choosing a high-quality observation instrument is only the first step; designing data-collection procedures for that instrument is the real work states face over the next several years. Decisions abound: how many observations should be done for each teacher? Who should do those observations—principals, master evaluators, or peers? How should raters be trained, and what level of accuracy should be required before they begin classroom observations? How should states report the validity and reliability of teacher scores? Do validity and reliability vary according to the subject or grade being observed or the qualifications of the observer?

While these questions may seem arcane, their answers determine the characteristics of teacher scores and thus the kinds of decisions evaluation data can influence. Take the decision about raters as an example. The most common teacher observation arrangement is for principals to evaluate their faculty absent input from outside raters and peers. Benefits to codifying this arrangement in new practice include refocusing principal efforts on instructional leadership and recognizing that principals' knowledge of school context may improve scoring accuracy.

The costs, however, are steep. Principals vary widely in their ability to assess instruction quality; even with training, it is not likely that all will be able to tell good from poor instruction in multiple content areas. The history of principal-based classroom evaluation also suggests many bow to the "culture of nice" prevalent in schools by rating roughly 97 percent of teachers as satisfactory or above; if this continues, it will entirely defeat the purpose of system redesign.

Most important, however, is that when teachers are rated by only their building principal, it is impossible to statistically separate principal bias (an overall tendency towards harshness, for instance) and teaching quality. This means that any comparisons of teachers based on principal ratings hold only within schools—that is, the district cannot assume two teachers with the same scores in different schools provide an equivalent level of instruction because the same principal did not score both of them.

Schools must account for this when making personnel decisions, basing terminations on comparisons within, not between, schools. A better system ensures that teacher scores reflect an average of multiple observers' impressions, a situation that can be easily constructed

using master or peer evaluators from either within the district or outside.

Decisions about the number and timing of observations are equally consequential for the new observation systems. Too few observations—or observations clustered too close together—may produce misleading teacher scores, analogous to determining a fourth grade student's overall proficiency based on a handful of test items on adding fractions. Too many observations and resources are wasted. A method known as a generalizability/decision-type study can provide information on the necessary number and type of observations by examining the multiple influences on score reliability. But evidence shows that few states have gone this route. In fact, our informal poll suggests states are making ad hoc decisions about the number of observations per teacher; one state intends to require four observations per year for tenured teachers, while another state's new legislation requires only one observation per year for tenured teachers. None reported plans to conduct any kind of generalizability study, instead relying on more primitive statistics such as percent agreement between raters for quality-control purposes.

The training and certification of raters in these new observation systems will also require an enormous investment. The process consists of three general parts:

1. The training modules themselves must have example of teaching—usually videos of instruction—for trainees to watch, rate, and learn from. Instrument developers must collect and catalog these examples so that experts on the rating instrument can master score them before they are assembled into training modules.
2. A certification test must be compiled to make sure raters are accurate.
3. Additional examples of teaching are needed for raters to practice before and after certification. In addition, raters must calibrate over the course of the school year to make sure they are on target.

By now, the picture is clear: the process of building the data-collection component of an observation system is slow and laborious and requires a considerable amount of specialized expertise. The decisions that can be made using the data also depend highly on the characteristics of the observation system.

Customize to maximize decision-making power. In the new teacher evaluation system we are proposing, teachers will fall into two groups: a large group with no real risk of termination or a similar outcome (such as a change in teaching status or assignment), and a much smaller group of teachers who do face such consequences. The latter group may include teachers without tenure, with low value-added scores, or with poor ratings based on previous observations.

In light of this reality, states should customize their teacher evaluation systems to deliver high accuracy and validity for the small set of high-stakes decisions while providing more general information on a looser schedule to the broader population of teachers. In practice, this means concentrating resources on at-risk teachers—for instance, conducting six observations instead of three, having each lesson rated by two observers, or even video-recording observations for scoring by external raters. Again, generalizability studies can help determine how many lessons and raters are necessary to meet a target level of reliability in teacher scores.

Align, align, align. Observers have long noted the difficulty of achieving reform when districts send multiple, conflicting messages to practitioners. New teacher evaluation standards and procedures increase the risk of this occurring. For instance, observation instruments may prioritize features of instruction not supported by curriculum materials; advice from external professional developers may directly contradict elements of new teaching standards; and additional emphasis on standardized testing through the use of value-added scores may conflict with a focus on teaching to develop competent students. An additional complicating factor in coordinating reform efforts is that in many districts, the teacher evaluation office is separate from the curriculum and instruction staff. Policymakers should carefully cross-walk elements of any new system with existing curriculum, professional development, standards, and assessments within the district.

Embed learning opportunities in the system. If successful, teacher evaluation reform would enable administrators to provide substantive, tailored feedback to individual teachers. To produce real improvements in teaching, such feedback would need to be coupled with opportunities to develop in target areas through mentoring, professional development, or other means. One route would be to develop professional development

specifically around evaluation criteria that prove difficult for teachers to master. Another would be to tightly couple existing professional development options to the action plans that result from teacher observations. Either way, we argue that the reform of the teacher evaluation system will see its chief successes not through carrots and sticks, but through providing teachers with information about their performance and means for improvement.

States should customize their teacher evaluation systems to deliver high accuracy and validity for the small set of high-stakes decisions while providing more general information on a looser schedule to the broader population of teachers.

Conclusion

It remains to be seen whether the sea change in teacher evaluation will provoke real reform or become just another layer of bureaucracy. Either way, differences among teachers are real, whether measured observationally or by student gains on assessments. Investing in systems to accurately measure this variation will help states and districts make smarter decisions about recruiting, hiring, and granting tenure to new teachers, as well as developing compensation strategies, career ladders, and professional growth plans for experienced teachers.

Several states and districts can serve as examples in this endeavor. Cincinnati, Ohio; the District of Columbia; Montgomery County, Maryland; and Tennessee have all revised and implemented more stringent teacher evaluation strategies: Cincinnati and Montgomery County over several years, and the District of Columbia and Tennessee much more quickly. These varied approaches are worth examining in detail. Examples of our recommendations, particularly in terms of alignment, can be found in these systems of teacher evaluation. Fundamentally, all have attempted to align their teacher evaluation system with broader goals for their students and teachers.

When reviewing or designing an evaluation system, look for multiple measures of teaching effectiveness, each

of which provides added information for school and district administrators, as well as teachers, to differentiate instructional practice and results for students. The observation system is most important, because if designed and implemented well, it will provide teachers the information they need to improve. Value-added scores or other measures of student achievement attributable to individual teachers also provide important information, but we argue that they are best used to direct scarce resources to more carefully review the practice of teachers identified as particularly high or low performing.

Evaluation systems need not be one size fits all. Customization can put greater focus and resources on new teachers or those flagged as weak in one or more areas and more efficiently improve the teaching corps. We also urge states, districts, and other stakeholders to use available research to guide the design of their systems and to build systems that can adapt to new and better evidence and lessons learned through early implementation.

Notes

1. Bill Turque, "More Than 200 D.C. Teachers Fired," DC Schools Insider, July 15, 2011, www.washingtonpost.com/blogs/dc-schools-insider/post/more-than-200-dc-teachers-fired/2011/07/15/gIQADnTLGI_blog.html (accessed October 18, 2011).
2. Eric A. Hanushek, "The Single Salary Schedule and Other Issues of Teacher Pay," *Peabody Journal of Education* 82, no. 4 (2007): 574–86; William L. Sanders and June C. Rivers, *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement* (Knoxville: Tennessee Value-Added Research and Assessment Center, 1996); John Schacter and Yeow Meng Thum, "Paying for High- and Low-Quality Teaching," *Economics of Education Review* 23, no. 4 (2004): 411–30.
3. Audrey Amrein-Beardsley, "Methodological Concerns about the Education Value-Added Assessment System," *Educational Researcher* 37, no. 2 (2008): 65–75; Heather C. Hill, Laura Kapitula, and Kristin Umland, "A Validity Argument Approach to Evaluating Teacher Value-Added Scores," *American Educational Research Journal* 48 (June 2011): 794–831; Cory Koedel and Julian R. Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy* 6, no. 1 (2011): 18–42; Daniel Koretz, "A Measured Approach: Value-Added Models Are a Promising Improvement, but No One Measure Can Evaluate Teacher Performance," *American Educator* (Fall 2008): 18–27, 39; Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125, no. 1 (2010): 175–214.
4. Michael Winerip, "Evaluating New York Teachers, Perhaps the Numbers Do Lie," *New York Times*, March 6, 2011.
5. Pamela Grossman et al., "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores" (working paper no. w16015, National Bureau of Economic Research, Cambridge, MA, 2010); Robert C. Pianta et al., "Classroom Effects on Children's Achievement Trajectories in Elementary School," *American Educational Research Journal* 45, no. 2 (2008): 365–87.
6. See the Measures of Effective Teaching Project website, www.metproject.org (accessed October 18, 2011). This project is capturing classroom observations of more than 3,000 teachers and using five observation instruments to measure teacher practice. The findings show how these measures relate to one another as well as to measures of student achievement.
7. We conducted an online poll in April 2011 of eight Race to the Top states and one nonparticipating state, asking them about their plans for teacher evaluation reforms.