The Future of Teacher Evaluation

Passage of the Every Student Succeeds Act (ESSA) in December 2015 signaled a new era for teacher evaluation reform. Under ESSA, states and districts have greater autonomy to design and implement teacher evaluation systems independent of federal influence. This new flexibility brings with it new responsibilities and challenges for states and districts as they consider ways to improve their current teacher evaluation systems—which teacher performance measures to choose, how to weight them, how to combine them, and where to set the rating thresholds that determine teacher effectiveness.

In this article, I identify these key challenges and opportunities and pose a series of questions for policymakers and school leaders to consider as they undertake revisions to their existing evaluation systems. These queries should inform state and district leaders on the issues involved in redesigning teacher evaluation systems and the potential consequences of their decisions.

Recent Revisions to Teacher Evaluation

Between the 2009–10 and 2015–16 school years, 88 percent of states and 89 percent of the largest 25 districts and the District of Columbia redesigned and implemented new teacher evaluation systems. These systems incorporated multiple measures of teacher performance—most prominently, student test-score data as value-added measures (VAMs) and more rigorous (and more frequent) classroom observation of teacher practice. The use of student test-score data and classroom observations was largely standardized in response to the demands of the Race to the Top (RTTT) federal grant program and No Child Left Behind state waivers, which required states and districts to use multiple measures of teacher performance (and multiple teacher ratings categories) in an effort to better identify and differentiate teacher effectiveness.

As state and district policymakers consider further revisions and potentially systematic changes to these new evaluation systems, key issues must be addressed:

- Which measures of teacher performance should be used to measure overall teacher effectiveness?
- What weight should be assigned to each teacher performance measure?
- How should a teacher’s summative evaluation score be constructed from multiple teacher performance measures?
- How should a teacher’s summative evaluation rating be determined, including the number of rating categories and percentage of points necessary (i.e., the rating threshold) for a teacher to earn a particular summative evaluation rating?

There are two important considerations implicit in these decisions. First, does a district have the human capital necessary to implement, for example, rigorous observations of classroom practice for all teachers, including in-class observation and pre- and post-observation conferences between teachers and their evaluators? Second, to what extent do different measures of teacher performance inform the underlying goals of teacher evaluation: providing teachers and school leaders with information to improve teacher instructional practice and differentiating teacher effectiveness for the purposes of high-stakes teacher personnel decisions (i.e., retention, promotion, dismissal)?

While a growing body of evidence-based research can guide state and district policymakers’ decisions, designing teacher evaluation systems also entails numerous value judgments and often a high degree of statistical and technical sophistication.

Several recent studies have found that state education agencies (SEAs) have limited technical expertise, which has

ESSA provides fresh impetus to amend recently minted teacher evaluation systems or revise them altogether, armed with this practical guidance for policymakers and school leaders.

by Matthew P. Steinberg
Reliability and Validity

As with any measure of performance, policymakers must be aware that measures of teacher effectiveness may be subject to reliability and validity concerns. Reliability relates to a performance measure’s variability; validity relates to a measure’s accuracy. For example, significant variability in a measure of individual teacher effectiveness—either from class to class or year to year—reflects a relatively unreliable (i.e., inconsistent) assessment of teacher effectiveness. A measure that provides inaccurate information about an individual teacher’s effectiveness—due to, for example, the nonrandom ways in which students are assigned to a teacher’s classroom or the nonrandom sorting of teachers across schools—reflects a relatively invalid (i.e., biased) assessment of teacher effectiveness. A number of empirical studies of teacher evaluation systems have examined the reliability and validity of measures of teacher effectiveness, including VAMs, classroom observation scores, and student surveys.

While concerns about the reliability and validity of teacher performance measures have been widely acknowledged in both academic and policy research, this should not mean that evaluation systems abandon these measures entirely. Instead, policymakers should consider ways to incorporate multiple measures of teacher performance into teacher evaluation ratings and, where necessary, make adjustments to these scores to reflect the fact that teachers and students are not randomly distributed across schools and classrooms.

What Measures of Teacher Effectiveness Should You Use?

Current teacher evaluation systems comprise one or more measures of teacher effectiveness, with each capturing a dimension of teacher performance. State and district administrators should consider the efficacy of each measure in identifying teacher effectiveness (to serve the accountability function of teacher evaluation), its ability to drive instructional improvement, and its fiscal and human capital requirements for full implementation.

Classroom Observation

Observations of teacher instruction are designed to provide formative feedback that teachers can use to improve their instructional practice as well as to carefully evaluate instructional performance. The observation process, in addition to the classroom observations themselves, often includes pre- and post-observation conferences between the teacher and the observer. Evidence suggests that rigorous classroom observation can be an effective means of improving student performance. Of the states and 25 largest districts that have implemented reforms since 2009, all have incorporated more rigorous, standards-based classroom observations into their summative teacher evaluation ratings.

What observation rubric will you use? Of the many observation rubrics available, the Danielson Framework for Teaching (FFT) has been the most widely adopted in newly implemented teacher evaluation systems. While the FFT has been shown to provide a complementary measure of teacher effectiveness (along with value-added measures, or VAM, and student surveys), the FFT by itself is limited in its ability to identify effective teachers. Further, there is growing evidence that measures of teacher
practice based on observation rubrics such as the FFT may be biased by the incoming achievement of a teacher's students. That is, teachers' observation scores tend to be higher when their students enter class having performed better on state standardized assessments in the preceding academic year. Other popular observation rubrics are the Classroom Assessment Scoring System (CLASS) and, for English language arts instruction, the Protocol for Language Arts Teaching Observations (PLATO). In addition to selecting a rubric, system designers must decide whether to use the full protocol or a subset of the protocol's domains.

**How many observations will you conduct?**
Administrators need to decide on the number of formal and informal classroom observations for both beginning (i.e., teachers in their first three years in the classroom) and career teachers. Of the states and largest districts that have implemented reforms since 2009, the largest districts conduct more observations than do states. The largest districts require, on average, 5.5 formal and informal observations for beginning teachers and 4.5 for career teachers. States require, on average, 4 observations for both beginning and career teachers. In addition to requiring more observations per teacher, districts are also more likely to require postobservation conferences between teachers and observers; this aspect of teacher observation can be crucial to teacher improvement.

**Who is responsible for conducting classroom observations and rating teacher performance?**
Will this task fall primarily to the principal, assistant principal, and/or a master teacher? Will evaluators receive the training necessary to conduct rigorous classroom observation?

---

**46 States Have Newly Redesigned Teacher Evaluation Systems Since 2011**

![Map of 46 states with new teacher evaluation systems since 2011]

*Source: Map data based on Figure 2 in M. Steinberg and M. Donaldson, “The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era,” Education Finance and Policy 11, no. 3 (2016).*
Policymakers must give careful consideration to the extent and quality of training provided to evaluators, since evaluator training is a key lever for improving teacher practice and student achievement.9

Do your school leaders have the capacity to conduct rigorous observations? Given that multiple classroom observations are required for each teacher, do districts and schools have the capacity to conduct all of the required observations? The observation process is time intensive, requiring evaluators to conduct not just multiple annual observations per teacher but also pre- and post conferences. Taken together, this can add up to 10 to 12 hours per teacher. In an era when educational leaders already say they don’t have enough hours in the day, administrators must consider how much time they can realistically devote to teacher observations. This is a particularly critical question for large urban districts, which tend to be resource constrained and yet require school leaders to conduct more observations than their suburban and rural counterparts. Further, since much of the efficacy of classroom observations for improving teacher practice and student achievement rests in formative and summative teacher-evaluator conferences; the capacity to conduct them is critical.

Value-Added Measures

VAMs measure teacher performance based on student achievement data, often from state-administered accountability exams. Only 30 percent of teachers nationwide teach subjects or grades for which a VAM score is available.10 Moreover, because VAM scores are provided to teachers after the conclusion of the school year and offer no guidance for how teachers may improve, their value as supports to improve instructional practice is limited. Eighty percent of both states and the largest school districts that have implemented new evaluation systems since 2009 use one or more measures of teacher performance based on student test score data (including VAM and student growth percentiles).11

Teacher evaluation reforms in the largest districts tend to rely on VAMs, while state evaluation systems tend to use student growth percentiles.12

What degree of uncertainty and variability are you comfortable with? Researchers have found substantial year-to-year variability in teachers’ individual VAM scores depending on the nature of the student assessment being used.13 While some have argued that this variability calls into question the ability of VAMs to identify effective teachers, others have found that students taught by higher-VAM teachers have better long-term educational and labor market outcomes.14

Do your districts have the technical capacity to develop VAMs? Constructing these measures requires statistical and technical sophistication. Administrators must assess whether they have the technical capacity to build and use these measures. In the absence of such capacity, external resources (such as university researchers and policy research organizations) increasingly support states and districts in constructing VAM scores for their teachers.

Student Learning Objectives

SLOs provide measures of student growth in grades and subjects that state accountability exams do not cover. Growth is measured based on subject- and grade-specific learning goals established either by state or national standards or by objectives that teachers or districts establish. Teachers are evaluated based on the percentage of students in their classes who meet these goals. Fifty-two percent of states and 39 percent of the largest districts that have implemented reforms since 2009 use SLOs in their teacher evaluation systems.

What assessments will you use? A student’s progress toward meeting SLOs may be evaluated using locally developed assessments or off-the-shelf tests. Districts must select SLOs for each grade and subject that the state’s accountability exam does not test. Further, the extent to which teachers and school leaders are involved in selecting the assessments incorporated into SLOs is an important consideration when designing evaluation systems.

Student Surveys

Student surveys measure teacher performance based on students’ perceptions. Seventeen percent of both states and the largest districts use student surveys as part of their teacher evaluation systems.16
What survey will you use? One widely adopted survey is the Tripod survey. Developed by Ron Ferguson, it is organized around seven domains of a teacher’s classroom practice. Students respond to 34 items on a five-point Likert scale ranging from totally true to totally untrue. As previously described, a composite measure of teacher performance—including student surveys, classroom observations, and VAM scores—can reliably identify effective teachers. However, there is limited evidence on the extent to which survey scores alone can distinguish teacher effectiveness and provide meaningful feedback to improve instructional performance. Nonetheless, students are key stakeholders in the process, and policymakers should consider not just whether to include such surveys into teacher evaluations but also the weight to give to surveys based on student feedback.

Schoolwide Achievement

Measures of schoolwide performance have increasingly been incorporated into newly implemented teacher evaluation systems. Thirty percent of states and 26 percent of districts use schoolwide achievement as a measure of teacher performance.

What aspects of school performance will you incorporate into schoolwide achievement measures? States and districts may choose to include different aspects of school performance into measures of schoolwide achievement. They may include measures of student achievement based on state accountability exams, indicators of the extent to which academic achievement gaps by racial/ethnic subgroups and socioeconomic status are closing, graduation rates, attendance rates, Advanced Placement or International Baccalaureate participation, and SAT participation, among others.

What might be the consequences of using schoolwide performance measures for individual teachers? Among states and districts that have recently implemented new teacher evaluation systems, an average of 4 percent of the summative evaluation score for a tested teacher (i.e., those with VAM scores) depends on schoolwide achievement. However, this percent increases to 7 percent (at the state level) and 12 percent (at the district level) for teachers teaching in nontested grades/subjects. Notably, schoolwide achievement may have little—if anything—to do with a nontested teacher’s individual performance and may therefore have important consequences for teacher retention. For example, economically disadvantaged schools may find it more difficult to retain high-performing teachers who teach in nontested grades/subjects, especially since a significant portion of their summative evaluation scores depend on the performance of students they do not teach. As a result, the ratings for highly effective, nontested teachers in large urban districts may reflect much poorer quality instruction than their own and may lead these teachers to look for teaching opportunities in higher-performing schools and districts.

Calculating a Teacher’s Summative Performance Score

When used in combination, classroom observations, VAMs, and student surveys are better able to identify and differentiate teacher effectiveness. As administrators reassess their teacher evaluation systems in light of the greater flexibility ESSA provides, they will want to consider how the overall teacher performance score is created from individual performance measures.

What nominal weights should you assign to each measure? Teacher evaluation systems may assign any feasible set of weights to the performance measures incorporated into a teacher’s evaluation, as long as the weights sum to 100 percent. However, these weights play a critical role in determining the proportion of teachers deemed to be professionally proficient. For example, when zero percent of a teacher’s summative evaluation score depends on classroom observations, the share of teachers deemed proficient (based on simulation analyses across eight large urban districts) ranges from 2 percent to 47 percent. In contrast, when 100 percent of a teacher’s summative evaluation score is based on classroom observations, the share of teachers deemed proficient ranges from 27 percent to 94 percent. Notably, despite the public attention on using student test-score data to evaluate teacher performance, student achievement measures (i.e., VAM scores and student growth percentiles) play a limited role...
in the evaluation of a typical teacher nationwide, constituting no more than 13 percent of a teacher’s summative evaluation rating.23

How do you want to treat teachers who have different measures available to them? Do you want to incorporate (and weight) different teacher performance measures—such as SLOs and schoolwide achievement data—into the summative evaluation score for teachers who do not have individual VAM scores? Doing so poses its own set of questions, such as how to make comparisons across teachers if different measures are contributing to their final evaluation score. And the greater reliance on schoolwide achievement data means that a larger share of a teacher’s score will be based on the achievement of students with whom she has little or no instructional contact.

Translating Individual Scores into Ratings

Translating a teacher’s summative performance score into a summative evaluation rating for the purposes of comparing teachers and making high-stakes personnel decisions requires some important decisions. Until recently, there has been little evidence to inform policymakers about how such design decisions determine the distribution of teacher ratings and the proportion of teachers deemed professionally proficient. This lack of guidance is problematic, particularly as these summative ratings are increasingly used to inform personnel decisions. A simulation-based analysis that I recently conducted with Matthew Kraft produced some estimates that can help to answer the following questions.24

How will you differentiate teacher ratings?

In other words, what thresholds will you use to sort teachers into categories of effectiveness? Following the federal RTTT directives, states and districts incorporate multiple categories. In most systems, absolute thresholds are used to determine a teacher’s summative evaluation rating. A teacher is placed into one of multiple (often four) rating categories based on the percentage of total evaluation points he or she earns. In evaluation systems that use absolute thresholds, all teachers may be rated in the highest category; such a system has limited ability to differentiate teacher effectiveness. Other districts impose a distribution of teacher effectiveness. In such a norm-referenced approach, the district assigns summative evaluation ratings to teachers based on a desired distribution of teacher effectiveness. For example, Dallas Independent School District assigns 40 percent of teachers to receive one of the lowest two ratings (Unsatisfactory or Progressing) and 60 percent of teachers to receive one of the highest two ratings (Proficient or Exemplary).25

Notably, among systems that use absolute thresholds, there is wide variation in the share of evaluation points a teacher must earn in order to be rated in one of the top two rating categories for which teachers are considered professionally proficient. As a result, the proportion of teachers deemed professionally proficient depends critically on where districts and states set thresholds. For example, a teacher who earns 60 percent of all possible summative evaluation points would receive the lowest possible rating in New York City; the second-lowest rating in Chicago, Denver, and Miami-Dade; but a proficient rating in Clark County, Fairfax County, Gwinnett County, and Philadelphia.26

Some Final Considerations

Do you want to account for contextual factors that influence teachers’ scores? In most systems, an overwhelming majority of teacher summative scores are based on classroom observations. But teachers who are assigned lower performing students receive lower observation scores, independent of the teachers’ ability or underlying effectiveness.27 To what extent might your evaluation system account for the contextual factors that bias measures of teacher performance? Such choices will affect cross-teacher comparisons. System designers must carefully consider whether the type of statistical adjustments used in the construction of VAMs should be used with measures of teacher performance based on classroom observation scores.

What effect will your decisions have on teacher motivation? State and district administrators must keep in mind how the selection of measures used to evaluate performance might motivate or discourage teachers. Will disproportionately weighting VAMs and SLOs...
in accountability systems dishearten teachers? Will teacher buy-in be greater in systems that incorporate and weight measures that focus on formative development, such as observations that entail pre- and postconferences? Do VAMs—which tend to be viewed by teachers as opaque, both in their construction and their value to teachers in terms of instructional improvement—discourage teachers from teaching in tested grades and subjects? It remains an open question whether the measures that states and districts use motivate teachers to improve or compel them to transfer out of grades or teaching assignments.

Looking Ahead

A growing body of evidence suggests that recent teacher evaluation reforms have improved teaching and learning and helped identify and remove the lowest-performing teachers. With the additional autonomy provided under ESSA, state and district administrators can further enhance their teacher evaluation systems—and tailor those systems to their schools’ and districts’ unique needs, values, and capacities. This article should provide foundational guidance to state policymakers and administrators as they review and consider revising their existing systems.

5Steinberg and Donaldson, “The New Educational Accountability.”
8Steinberg and Donaldson, “The New Educational Accountability.”
9Steinberg and Sartain, “Does Teacher Evaluation Improve School Performance?”
11Student growth percentiles and VAMs both measure teacher performance based on student achievement data. However, student growth percentiles do not statistically account for student or class characteristics in the way that VAMs do.
12Steinberg and Donaldson, “The New Educational Accountability.”
15Steinberg and Donaldson, “The New Educational Accountability.”
16Ibid.
17Kane et al., “Have We Identified Effective Teachers?”
18Steinberg and Donaldson, “The New Educational Accountability.”
19Ibid.
20Ibid.
21Kane et al., “Have We Identified Effective Teachers?”
23Steinberg and Donaldson, “The New Educational Accountability.”
25Ibid.
26Ibid.
27Steinberg and Garrett, “Classroom Composition.”