Supporting Effective Teacher Evaluation

In almost all states, education policy focuses on teacher evaluation in a bid to assess how well teachers are handling the substantial demands placed upon them. Teacher evaluations are intended to improve instruction in the classroom, and yet many state policies preclude the possibility that a teacher is both a productive employee and in need of improvement. Yet this situation characterizes most classrooms today. States that fail to address this disconnect between policy and practice will also fail to mount any meaningful improvement in instruction.

Teacher evaluation varies greatly from state to state and subsumes a range of policies, practices, and assumptions. To understand how systems of evaluation are changing and where they fall short, it is essential to understand some common drivers behind changes in the states’ systems of evaluation:

- National and international test results, reports from employers and higher...
Research has identified teachers as the most important school factor related to student achievement.² While research also finds that factors outside the school are more important in determining student outcomes, this distinction at times disappears altogether from political rhetoric.

Parents and policymakers have expressed deep frustration with the educational system’s ability to identify and take action on poorly performing teachers. The evaluation process typically fails to make much differentiation among teachers, leading to distrust of the status quo.³

New statistical methods known as value-added modeling (VAM) purport to link student learning outcomes as measured by standardized tests to specific teachers.⁴ These models are designed to overcome the problem of crediting teachers as being more or less effective based on achievement levels (which are highly related to socioeconomic status) of the students they teach. The models take into account students’ prior test scores and attribute the changes in year-to-year performance to a specific teacher.

There is a need to evaluate teachers’ contributions to student learning for those who teach in grades and subjects that are not subject to annual standardized testing.

Education policymakers increasingly understand the need to include measures of teacher practice, most often through classroom observation, as part of an evaluation system.

**Components of Teacher Evaluation Systems**

These forces have led to the redesign of state teacher evaluation systems. Federal programs such as Race to the Top, School Improvement Grants, Teacher Incentive Fund, and the Elementary and Secondary Education Act have further incentivized changes in teacher evaluation. Despite the variety across states and often across districts within states, all systems include a student growth and a teacher practice component that are then combined to determine an overall teacher evaluation score (figure 1).

---

*Figure 1. Teacher Evaluation Scoring*
In most states, different combinations of the following measures, along with different aggregation schemes, are used to determine an overall student growth measure.

**Standardized growth measures.** Virtually all states include a measure of student growth based on standardized achievement tests. However, some states use value-added modeling (VAM), while others use student growth models (SGMs). The specifics of the VAM and SGM models also vary across states. These growth measures are typically available for a minority of teachers in the state as a result of state testing requirements for specific grades and subjects.

**Student learning objectives (SLOs).** Used by many states, SLOs are locally determined measures of teacher effectiveness in which measurable targets for student achievement are set following an analysis of baseline data. The extent to which targets are met is then used to evaluate the teacher based on the degree that her students attain the goals. Every SLO has three primary components: the population of students it covers, the target for student achievement through the year, and the assessments that are used to evaluate the target. Specific features of the SLO process vary substantially, including the extent to which targets are common across groups of teachers, the nature of the assessments (e.g., commercially available, developed to be used across groups of teachers, or unique to particular teachers), and the extent to which some or all of a teacher’s students are included.

**School or gradewide measures.** Some states also include a school or gradewide growth measure to signal the collective responsibility that teachers have to ensure student progress within the school.

**Proficiency measures.** Some states also include a proficiency measure for some group of students (e.g., average test scores for the school) as part of the student growth measure. This is intended to recognize absolute academic performance in addition to the idea of growth embedded in the other constituent measures.

---


*Many states have analogous measures but use different terms: New Jersey uses student growth objective; North Carolina uses measures of student learning; Massachusetts uses district-determined measures; and Washington, D.C., uses teacher-assessed student achievement.*

State policymakers have deployed a variety of measures that fall within each component (see boxes 1 and 2).

### Relevant Distinctions for Practice

Describing the components of teacher evaluation provides only a superficial understanding of the status of these systems and the promise they have for improving education. To more fully appreciate the potential of teacher evaluation, state boards of education and state education agencies must consider the policies, practices, assumptions, and lenses that define their systems and influence their implementation, quality, and potential outcomes.

The distinctions drawn here are not intended to be exhaustive. Nor are they independent. Decisions and actions taken for particular aspects of the system have direct implications for other aspects and at times can work at cross-purposes.

### Purposes and Policies

Most state evaluation systems have two explicit goals. One is to evaluate teachers for purposes of making employment-related decisions. On the basis of poor evaluations, teachers may be placed on probationary status or terminated. Conversely, strong evaluations may in some states lead to increased pay and reduced demands in future evaluations. The second goal is to improve classroom instruction. The key question for state and local policymakers is how these two goals can coexist productively.

Why is this so? To understand this disconnect, it is important to first note that recent research, based on observations or other measures and studies of classroom practice, judges large numbers of classroom interactions to be in need of improvement. Particularly on dimensions associated with instructional rigor,
the majority of classroom ratings fall somewhere below the midpoint on the measurement scales (e.g., a 2 on a 4-point scale). Such scores are likely to be termed either “basic” or “needing improvement,” depending on the scale used. Thus the research message is consistent: The majority of US students need to improve performance, and the majority of US classrooms need better instruction.

A reasonable person reading such findings would conclude that the quality of teaching across a broad proportion of classrooms should be improved. However, many state policies convey a very different message—that being assigned a rating of “needs improvement” puts someone on probationary status, potentially a precursor to being fired. Where policies treat “needs improvement” as an actionable status, the likelihood of improving educational performance is utterly compromised.

If saying someone needs improvement carries with it such dire disciplinary consequences, administrators will justifiably designate as needing improvement only those teachers who they consider to be weak employees. In such a system, “needs improvement” actually is understood to mean that a teacher’s performance is unsatisfactory. Thus, while many state policies preclude the possibility that a teacher is both a productive employee and in need of improvement, available research indicates that this is exactly the sort of person who characterizes most of the profession. At the same time, their supervisors are classifying most teachers as proficient/effective or highly proficient/effective.

The feedback from measures of teacher practice is similarly compromised. Teachers receive inflated observation scores that do not accurately describe the quality of their practice, making it difficult to provide effective feedback. In the majority of teachers’ classrooms, for example, the quality of classroom discourse is limited (a 2 on a 4-point scale). Teachers still control most of the classroom talk, and in many classrooms opportunities for students to share and exchange thinking are limited. Yet the vast majority of teachers will receive scores of 3 or higher, which are associated with a qualitative description in various protocols as classrooms that are rich in the exchange of ideas among and between teachers and students. It would be difficult to motivate teachers to change this aspect of teaching practice if they are receiving scores that convey high performance.

State boards of education should establish policies that support the goals of evaluation and instructional improvement. Forcing classifications that corrupt the meaning and use of measures designed to illuminate teaching will necessarily compromise initiatives designed to provide meaningful feedback and improve practice.

**Tenuous Link between Research and Practice**

A great deal of research has been used to inform evaluation systems’ overall design and the justification for them. Most prominent is the Measures of Effective Teaching Project. One of its most important findings (and those of related studies) has been the identification of sources of measurement error in order to ensure that judgments about teachers are as valid as possible.

First, classroom observation scores vary because of who observes, when they observe, and what kinds of lessons they observe. Any single observation by a single rater is a relatively unreliable measure. Thus research recommends that observers be highly trained, pass certification tests, and continue to have their performance checked and calibrated to ensure that they maintain a clear understanding of how to apply observation protocols. Further, it is recommended that every teacher be observed multiple times, by different observers, and with reliability checks to evaluate the continuing quality of the observations. To the extent possible, ratings should be a function of what is observed, not who is observing.

In the cauldron of practice, though, controls for measurement error are largely absent. Observers receive minimal training, and though they may have to pass a certification test, both training and certification is far less rigorous than that which observers experienced in these research studies. Additional postcertification calibration is rare, as is rating of a lesson by more than one observer to check on the reliability of observers. In some cases, the number of mandatory observations has been reduced. Thus, quality controls to ensure the quality of scores are in practice absent. Perhaps the most critical distinction between measures used in research and those used in practice is that in research, observers and evaluators have no relationship

---

Many states’ and districts’ use of evaluation for instructional improvement has been quite limited.
or knowledge of the people being observed and thus the researchers’ decisions have no bearing on the day-to-day and organizational impact of any evaluative judgments.

Given all the constraints and pressures on teachers and administrators, it is not possible for education systems to replicate these research efforts, but it might be possible to institute practices that provide meaningful information about the quality of evaluations. One can imagine periodic audit processes that add external, independent judgments against which those inside the system can calibrate their judgments and make modifications as appropriate.

**Policy Imperatives and the Distortion of Measurement**

Sometimes policies make sound measurement difficult. For example, one policy imperative—student growth—has been included as a central component of all teacher evaluations. The VAM research outlined earlier supports this policy. However, a relatively small proportion of teachers instruct students in the grades and subjects covered by annual standardized tests, so value-added estimates cannot be made for most teachers. For these teachers, policy requires that student growth be measured through a variety of locally determined processes, collectively known as student learning objectives (SLOs).

My colleagues and I have argued elsewhere that it is a mistake to construe SLOs as growth measures. Their various forms make them incomparable. When teachers and administrators treat SLOs as growth measures, they tend to focus on the attainment of somewhat arbitrary and problematic goals. For example, claiming that students should increase their scores from pre- to post-test by 20 percent is only meaningful if it is either clear how much most students improve in a year or it is clear what 20 percent means in terms of concepts or skills learned. In most cases, SLOs do neither. The goals risk being either trivial or so ambitious that they are unrealistic. Simply calling them growth measures does not make them so.

However, SLOs do have great potential if properly construed and supported as measures of teacher practice—how teachers establish important learning goals, how they develop assessments to measure that work, and how they consider student performance. These are central aspects of practice, and SLOs could be effective ways of both measuring and informing teacher practice. Indeed, there are successful models of SLOs being used in this way.

**Individual Differences versus Institutional Consistencies**

The way researchers frame their questions determines the results they find. In the case of research on teaching quality, an individual differences perspective has dominated it. Such studies seek to determine where individuals fall on a distribution of effectiveness and then to highlight those distinctions. This leads us to policies and practices that try to identify individuals at the tails of the distribution—particularly the tail inhabited by the weakest performers. But such a lens may preclude a grasp of the larger picture. When one takes an individual differences perspective, one inherently looks for differences. But when looking for commonalities, one is struck by the relative sameness of instruction within a school or across schools in a district. While there are certainly outliers, by and large classrooms do not differ very much on observational measures. An institutional focus would bring attention to the commonality of practice and direct policymaking toward actions that could improve instruction across the board: modified curricula, perhaps, or coordinated professional development.

An individual differences model alone would be reasonable if the US educational system were deemed high performing. In that case, an evaluation system that filtered out individuals who did not meet the standards of a high-performing system would make sense. However, in a system that could benefit from broad improvement, it makes more sense to focus on consistencies and institutional factors that influence teaching performance, while also taking action on the relatively small proportion of teachers designated as poorly performing.

**Implications for the Future**

Teacher evaluation will remain an entrenched feature of American schools. The challenge is to make evaluations as effective as possible in terms of improving instruction and student...
outcomes. This brief review suggests a few possible directions that state boards of education and state education agencies can pursue:

- **Do no harm.** Make sure that new policies do not subvert the intentions of the system one is trying to improve. One important way to gain insight into how the field is reacting to policy is to conduct occasional research studies that reveal what is working, or not working, as intended.

- **Recognize that evaluation processes are a set of tools to support professional judgments rather than a rigorous measurement process.** Ultimately, the effectiveness of the system will depend on the quality of evaluators’ judgments. Policies and processes should be set to monitor and improve the quality of those judgments.

- **Supplement individual evaluations with rigorous institutional evaluations.** If evaluations are going to improve education, then educators need to shift practice broadly and not simply identify and take action on the weakest. It is critical not to just compare individuals but to compare the status quo with where we all want our education systems to be. ■

---


