COMMENTARY

Measuring Teaching Effectiveness

By Linda Tyler

Across the nation, policymakers and education leaders share growing agreement that we must do a better job in measuring effective teaching and identifying effective teachers. And increasingly, there is a belief that measures of effective teaching should contribute to high-stakes decisions like pay for performance and tenure.

But people disagree about what it would take to identify effective teaching consistently and accurately. And many teachers worry that important contributions they make to schools and to the growth and well-being of students will be ignored.

Identifying good teaching doesn’t and shouldn’t have to be a matter of conjecture or opinion. We can build on solid research and create tools that accurately distinguish teaching that leads to student success from teaching that does not. And we can use multiple measures, including student-performance data, classroom observation, feedback from students, and other evidence, to provide well-rounded, fair, and valid input into important decisions.

For most of the high-stakes testing we do in education, we hold the instruments and supporting processes to high technical standards. We must do the same for measures of teacher effectiveness. To implement a robust and defensible system of performance measures, we will have to make sure that such measures are held to higher technical standards than many of today’s evaluation tools.

The vast majority of teacher-evaluation tools used today have not been demonstrated to measure what consistently leads to student learning. Most of them represent what particular groups of experts believe is central to good teaching, but very little evidence can be produced to show that these tools identify key components of practice that actually help
students learn. Gathering such evidence and validating the tools takes a focused and intensive effort.

Based on my work at the Educational Testing Service, I believe that assessments of teacher effectiveness used for high-stakes purposes should meet four key technical requirements:

**Generalization.** The results of the measure (or multiple measures) must be shown to adequately represent the scope and quality of a teacher’s performance. A 15-minute teaching segment is unlikely to convey in depth a teacher’s practice, but how much evidence, and what kind, is needed to make a valid judgment?

**Evaluation.** Scoring rubrics must be well aligned with what is measured, and the scoring itself must be accurate and consistent across time, across evaluators, and across the teachers being measured. This means that we will need to intensify training, calibration, and monitoring among those who observe and evaluate teaching in the classroom.

**Extrapolation.** We must be able to show that the performance results are a good gauge of our definition of teaching quality. Do high scores for teachers correlate to more or deeper learning for students?

**Implication.** The use of the performance results must be consistent with the original purpose of the assessment or be shown to be appropriate. An evaluation tool created to check for accomplished teaching, for example, would probably be inappropriate for evaluating the effectiveness of beginning teachers.

The right kind and combination of technically strong measures of teaching effectiveness will not only lead to better judgments about teachers, but also should provide feedback that most teachers don’t get today—feedback that could inform their professional growth. In fact, an ideal evaluation system would go hand in hand with teachers’ own charting of their course for development and growth, using evaluation and assessment feedback to help in that planning. Imagine an electronic portfolio with not just a teacher’s own choice of exemplary work and results, but also data from student assessments and feedback from evaluations, peer coaches, parents, and students.

"Many teachers worry that important contributions they make to schools and to the growth and well-being of students will be ignored."
The good news is that efforts are moving in the right direction, and with a clear sense of urgency, to expand what we know about the technical qualities of evaluation tools. The Bill & Melinda Gates Foundation is funding a number of projects to identify valid indicators of excellent teaching. These projects are examining the technical quality of several existing assessment instruments, and piloting early versions of new tools, from classroom evaluation tools, to pedagogical content-knowledge tests, to surveys of student perceptions. The data gathered on these tools will be compared with evidence of student outcomes, and combinations of measures will be simulated to determine which “multiple measures” might work best.

These studies are also incorporating innovative ways to gather teaching evidence, including new means of video capture, transfer, storage, and scoring. The first wave of data collection has begun, with initial results to be available this year.

The technical quality of measuring effectiveness, however, does not end with the instruments themselves or the means of collecting evidence. The quality must pervade how the measures are implemented, not just what measures are implemented. This means that classroom observation will require a substantial effort to provide adequate training for those who will evaluate, rigorous requirements to show that evaluators are applying scoring criteria consistently, and monitoring or quality-checking of scorers to make sure those judgments stay on track over time and in different classrooms. For teachers who do not teach in grades or subjects covered by state assessments or end-of-course tests, more work is needed in identifying, collecting, and evaluating alternative sources of evidence of student learning.

The bottom line is that we must do the work needed to ensure that measures of effectiveness are fair, rigorous, valid, and defensible, and that they result in feedback that teachers can apply to their professional growth. We owe this to teachers, and we owe it to students. The issues are complex, but not unsolvable. This won’t take a decade, but will take two or three years.

If we want to improve teaching and learning, the right combination of measures, including appropriate use of student-performance data, is a critical part of doing it.

*Linda Tyler is the vice president for teacher licensure and certification at the Educational Testing Service, a nonprofit educational measurement organization based in Princeton, N.J.*