Capturing the DIMENSIONS of Effective Teaching

When the world is in danger and it's time to summon the superheroes to save the day, my six-year-old son dives into his toy bin. Just like the comic-book authors, he emerges with a diverse team of superheroes, each with a *different* superpower. (I've noticed he never chooses three Supermen or four Spidermen, for instance.) One will have awesome physical strength but lack strategic vision; one will fly or run with superhuman speed but be impulsive and irresponsible; and another will lack strength and speed but make up for it with tactical genius (often combined with some dazzling ability, such as creating a force field or reading minds). The team always prevails, as its combined strengths compensate for the weaknesses of its members.

Student achievement gains, student surveys, and classroom observations

In the largest study of instructional practice ever undertaken, the Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) project is searching for tools to save the world from perfunctory teacher evaluations. In our first report (released in December 2010), we described the potential usefulness of student surveys for providing feedback to teachers. For our second report, the Educational Testing Service (ETS) scored 7,500 lesson videos for 1,333 teachers in six school districts using five different classroom-observation instruments. We compared those data against student achievement gains on state tests, gains on supplemental tests, and surveys from more than 44,500 students.

So far, the evidence reveals that my son's strategy when choosing a team of superheroes makes sense for teacher evaluation systems as well: rather than rely on any single indicator, schools should try to see effective teaching from multiple angles.

By THOMAS KANE

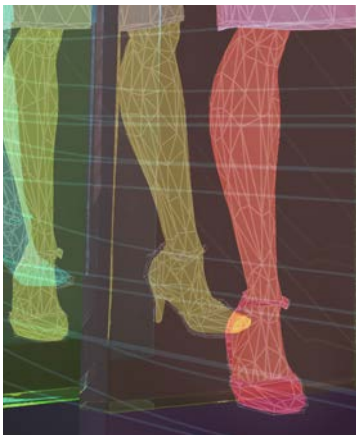
ILLUSTRATION / LONNIE BUSCH

Achievement Gains and Predictive Power

A teacher's track record of producing student achievement gains does one thing better than any other measure (even if it does so imperfectly): it signals whether a teacher is likely to achieve similar success with another group of students. Not surprisingly, this is particularly true when the outcomes are being measured with the same test. In comparison to classrooms of students elsewhere with similar baseline achievement and demographics, a teacher's achievement gain in one year is correlated at a rate of .48 in math and .36 in English language arts (ELA), with the average growth of students in another year. Such volatility notwithstanding, a track record of achievement gains is a more reliable predictor of the gains of future students than classroom observations or student surveys.

Unfortunately, state tests do not measure every outcome parents and taxpayers (and students) expect from schools, and cost is a factor in determining what gets measured. Given the higher cost of scoring constructed-response items, many states rely heavily on multiple-choice items to measure student achievement. The shallowness of the items on the test does not necessarily translate into shallow teaching. (For example, although spelling can be tested with low-cost items, a language teacher may find it useful to briefly summarize the reach of the Roman Empire while explaining the appearance of many Latin roots in the English language. A conceptual understanding can provide a framework for learning the fact-based knowledge examined on state tests.) In our study, the teachers with larger gains on low-cost state math tests also had students with larger gains on the Balanced Assessment in Mathematics, a more-expensive-to-score test designed to measure students' conceptual understanding of mathematics.

Our results did raise concerns about current state tests in English language arts, however. Current state ELA assessments overwhelmingly consist of short reading passages, followed by multiple-choice questions that probe reading comprehension. Teachers' average student-achievement gains based on such tests are more volatile from year to year (which translates to lower reliability) and are only weakly related to other measures, such as classroom observations and student surveys.



Teachers need to be able to see their own strengths and weaknesses clearly and recognize where they need to hone their skills.

We supplemented the state tests with an assessment requiring students to read a passage and then write short-answer responses to questions about the passage. The achievement gains based on that measure were more reliable measures of a teacher's practice (less variable across different classes taught by the same teacher) and were more closely related to other measures, such as classroom observations and student surveys. In order to provide clearer feedback on teacher effectiveness, states should hasten efforts to add writing prompts to their literacy assessments.

We expect schools to do more than raise achievement on tests, however. Parents hope their children will learn other skills that lead to success later in life, such as an ability to work in teams and persistence. Just because these skills are hard to measure and are not captured directly on any state test need not imply that effective teachers are ignoring them. Indeed, building student persistence may be an effective strategy for raising achievement on state tests. Recent evidence suggests that the teachers with larger student-achievement gains on state tests also seem to have students with greater long-term career success. As Raj Chetty, John Friedman, and Jonah Rockoff reported recently (see "Great Teaching," *research*, Summer 2012), being assigned to a teacher with a track record of student achievement gains is associated with higher earnings and rates of college going.

In sum, the "superpower" of the student achievement–gain, or growth, measure is its ability to "foresee" the achievement gains of future students and future earnings of students. But, like my son's flawed heroes, it also has drawbacks. One key weakness of the student achievement–gain measure is the limited number of grades and subjects for which assessment data are currently available. In many school districts, fewer than one-quarter of teachers work in grades and subjects where student achievement gains are tracked with state assessments.

In addition, student achievement gains provide few clues for what a teacher might do to improve her practice. A performance-evaluation system should support growth and development not just facilitate accountability. Teachers need to be able to see their own strengths and weaknesses clearly and recognize where they need to hone their skills. That is not information a value-added measure can provide.

Classroom Practice

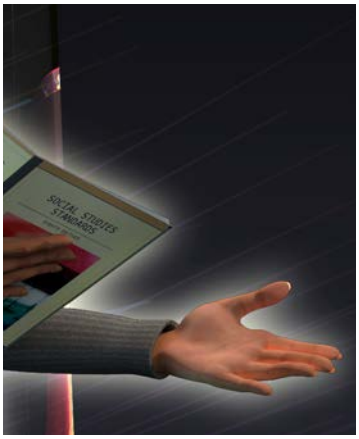
One way to develop such feedback is by means of classroom observation by a trained adult. Over the years, education researchers have proposed a number of instruments for assessing classroom instruction. To test these approaches, the Educational Testing Service trained more than 900 observers to score 7,500 lesson videos using different classroom-observation instruments. Depending on the instrument, observers received 17 to 25 hours of initial training. At the end of the training, observers were required to score a set of prescored videos. If the discrepancy between their scores and the master scores was too large, they were prevented from participating. (Across all the instruments, 23 percent of trained raters were disqualified because they could not apply the standards accurately.)

Every video was rated at least three times: once using the Framework for Teaching, developed by Charlotte Danielson; once using the Classroom Assessment Scoring System (CLASS), developed by Bob Pianta and Bridget Hamre at the University of Virginia; and a third time using a subject-specific instrument. The math lessons were scored using the Mathematical Quality of Instruction (MQI), developed by Heather Hill at Harvard. The ELA videos were scored on the Protocol for Language Arts Teacher Observation (PLATO), developed by Pam Grossman at Stanford. Finally, the National Math and Science Initiative scored a set of 1,000 math lessons, using the Uteach Observation Protocol.

I'm often asked, "Do you really think you can quantify the 'art' of teaching?" I argue that is not the right question. Of course, it is impossible to codify *all* the nuances that go into great teaching. But an instrument need not capture all the dimensions of great teaching in order to be useful. Each of the classroom-observation instruments proposes an incomplete but discrete set of competencies for effective teaching and provides a description of differing performance levels for each competency. The instruments' usefulness depends not on their completeness but on the demonstrated association between the few discrete competencies and student outcomes.

For example, one of the competencies highlighted by the Framework for Teaching is questioning skill. A teacher would receive an "unsatisfactory" score if she asked a series of yes/no questions, posed in rapid succession, to the same small group of students. A teacher would receive an "advanced" score on questioning skill if she asked students to explain their thinking, if the questions involved many students in class, and if the students began asking questions of each other. Depending on the instrument, observers tracked 6 to 22 different competencies, including "behavior management," "time management," and "engaging students in learning."

The goal of classroom observations is to help teachers improve practice, and thereby improve student outcomes. A classroom-observation system that bears no relationship to student outcomes will be of no use in improving them. As a result, we tested the relationship between classroom observations and a teacher's average student-achievement gains. All five of the instruments yielded scores that were related to student achievement gains, in the classroom of students where the teacher was observed as well as in other classrooms of students taught by the same teacher.



An instrument need not capture *all* the dimensions of great teaching in order to be useful.

In theory, classroom observations allow teachers to be more discerning about their own practice, and their improved practice will yield improved student outcomes. This is as yet a "potential superpower" of classroom observations, since there's not a lot of evidence that providing such feedback leads to improved student outcomes.

The poor track record of professional-development interventions provides ample reason for caution. Yet there is some reason for optimism. Eric Taylor and John Tyler report that midcareer teachers in Cincinnati saw significant improvements in student outcomes in the years during and after intensive observations (see "Can Teacher Evaluation Improve Teaching?" *research*, page 78). In fact, the gains in student outcomes were similar in magnitude to those seen during the first three years of teaching. It may be that professional growth must begin with an individualized (and honest) assessment of a teacher's strengths and weaknesses. We need better evidence in the coming years on the types of feedback and support that lead to improved student outcomes.

There are some downsides to classroom observations. First, if they are the sole basis for a teacher evaluation (as is true in many systems now), they may stifle innovation, forcing teachers to conform to particular notions of "effective practice." Second, each

of the instruments requires judgment on the part of observers. Even with trained raters, we saw considerable differences in rater scores on any given lesson. Moreover, possibly because different content requires teachers to exhibit different skills, a teacher's practice seems to vary from lesson to lesson. Even with trained raters, we had to score four lessons, each by a different observer, and average those scores to get a reliable measure of a teacher's practice. Given the high opportunity cost of a principal's time, or the salaries of professional peer observers, classroom observations are the costliest source of feedback.

Student Surveys

Student evaluations are ubiquitous in higher education, where they are often the only form of feedback on instruction. (Student achievement gains and classroom observations are rarely used at the college level.) The MET project investigated the usefulness of student evaluations in 4th-grade through 9th-grade classrooms.

To collect student feedback, the project administered the Tripod survey, developed by Ronald Ferguson at Harvard's Kennedy School of Government. Rather than being a popularity contest, the Tripod survey asks students to provide feedback on specific aspects of their classroom experiences. For example, students report their level of agreement to statements such as, "In this class, we learn to correct our mistakes," "Our class stays busy and does not waste time," and "Everybody knows what they should be doing and learning in this class." While administering the survey, we took steps to protect students' confidentiality, such as providing students with thick paper envelopes for submitting paper-based surveys or secure passwords to submit web-based surveys.

We learned several important lessons: First, students perceive clear differences among teachers. For example, in a quarter of classrooms, less than 36 percent of students agreed with the statement, "Our class stays busy and does not waste time." In another quarter of classrooms, more than 69 percent of students agreed.

Second, when teachers taught multiple sections of students, student feedback was often consistent. The between-classroom correlation in Tripod scores was .66. This is higher than we saw with the achievement gains measure. Attaining a comparable level of consistency with classroom observations required scoring four different lessons, each by a different observer. We had to average over multiple observations by multiple observers to generate reliable scores. Even if the typical student is less discerning than a trained adult, the ability to average over many students (rather than one or two adults), and having students experience 180 days of instruction (rather than observe two or three lessons), obviously improves reliability.

Third, the student responses were more correlated with teachers' student-achievement gains in math and ELA than the observation scores were. (Just as we did with classroom observations, to avoid generating a spurious correlation between student survey responses and achievement scores for the same group of students, we estimated

Combining Strengths (Figure 1)

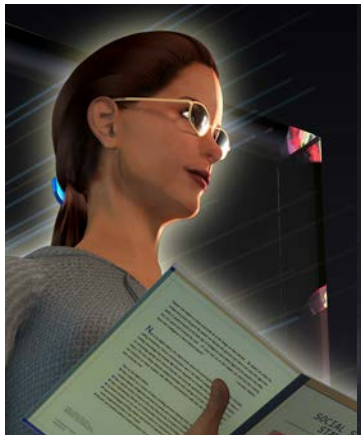
Achievement gains in one class are strong predictors of gains in another class taught by the same teacher; combining evidence from multiple sources yields a more reliable measure of teacher effectiveness.

Note: Predictive power is the unadjusted correlation with student achievement gains in another course section taught by the same teacher in the same year. Reliability is the correlation of the measure across course sections taught by the same teacher in the same year. The combined measures incorporate information from achievement gains, student surveys, and classroom observations. Criterion weights were generated by regressing achievement gains in one class on all three measures from another class taught by the same teacher. These findings pool elementary and middle school classrooms together. Future analyses will assess robustness to different statistical assumptions and will produce separate analyses for different grades.

SOURCE: Author's calculations

the correlation across different classrooms of students taught by the same teacher.) In other words, student responses were not only consistent across classrooms, they were predictive of student achievement gains across classrooms.

For those many states and districts that are struggling to find ways to measure performance in non-tested grades and subjects, well-designed student surveys should be an attractive option for supplementing classroom observations. They are also among the least costly of the measures.



Whenever one places all the stakes on any single measure, the risk of distortion and abuse goes up.

The Case for Multiple Measures

As with superheroes, all the measures are flawed in some way. Test-based student-achievement gains have predictive power but provide little insight into a teacher's particular strengths and weaknesses. Classroom observations require multiple observations by multiple observers in order to provide a reliable image of a teacher's practice. The student surveys, while being the most consistent of the three across different classrooms taught by the same teacher, were less predictive of student achievement gains than the achievement-gain measures themselves.

Fortunately, the evaluation methods are stronger as a team than as individuals. First, combining them generates less volatility from course section to section or year to year, and greater predictive power. Figure 1 compares the three different methods (classroom observations, student surveys, and student achievement gains) on reliability and predictive power. On the horizontal axis is the reliability of each method. (We report reliability as the correlation in scores from classroom to classroom taught by the same teacher.) On the vertical axis is predictive power, or correlation with a teacher's average student-achievement gain working with a different group of students in 2009–10. Both predictive power and reliability are desirable traits, so values in the upper-right-hand corner of the graph are more desirable. The student achievement–gain measure is most highly correlated with student achievement gains but has lower reliability than student surveys. Student surveys have the highest reliability but are less correlated with student achievement gains. Classroom observations, based on the Framework for Teaching, are less reliable and less correlated with achievement gains.

Figure 1 also reports two different combinations of the three measures: an "equally weighted" combination (standardizing each of the measures to have equal means and variances and then applying a weight of .33 to each) and a "criterion-weighted" combination. (To generate the weights, we regressed a teacher's average student-achievement gain in one class against the three different measures from another class, resulting in weights of .758, .200, and .042 on value-added, student survey, and classroom observation, respectively). The "criterion-weighted" measure offers more of the two desirable properties—predictive power and reliability—than any of the measures alone. (Even though classroom observations do not add much predictive power, it is hoped that classroom observations excel on a third dimension, not captured in the graph: the ability to diagnose specific strengths and weaknesses.) The next MET project report will explore weighting strategies in depth (see sidebar, page 40).

A second reason to combine the measures is to reduce the risk of unintended consequences, to lessen the likelihood of manipulation or "gaming." Whenever one places all the stakes on any single measure, the risk of distortion and abuse goes up. For instance, if all the weight were placed on student test scores, then the risk of narrowing of the curriculum or cheating would rise. If all the weight were placed on student surveys (as happens in higher education), then instructors would be tempted to pander to students and students might be more drawn to play pranks on their teachers. If all the weight were placed on classroom observations, then instructors would be tempted to go through the motions of effective practice on the day of an observation but not on other days.

The use of multiple measures not only spreads the risk but also provides opportunities to detect manipulation or gaming. For example, if a teacher is spending a disproportionate amount of class time drilling children for the state assessments, a school system can protect itself by adding a question on test-preparation activities to the student survey. If a teacher behaves unusually on the day of the observation, then the student surveys and achievement gains may tell a different story.

There is a third reason to collect multiple measures: conflicting messages from the multiple sources of information send a signal to supervisors that they should take a close look at what's going on in the classroom. Suppose a teacher is employ-

Upcoming MET Project Reports

In addition to the two reports released thus far, which are available at www.metproject.org, the MET project will be releasing two additional reports in early 2013. One will evaluate alternative approaches to weighting student achievement gains, classroom observations, student feedback, and the test of pedagogical content knowledge to develop a composite measure of effective teaching. We will explore a variety of rationales for combining measures (such as predicting underlying value-added on state tests, supplemental tests, and other student outcomes) and will describe the implications of each.

In the final report, we will address the issue of causality. The most vexing question we face is whether or not any of our results were biased by the exclusion of important student characteristics from the value-added models. Of course, there are an infinite number of additional student and peer characteristics, many of which are related to student achievement. The existence of these unmeasured determinants of achievement does not, by itself, imply bias, nor would it necessarily cause bias if teacher assignments are based partially on such factors. Rather, the question is whether or not such unmeasured traits are systematically related to the measures we use—classroom observations, student surveys, value-added estimates, and so forth.

Ultimately, the only way to resolve such questions is by randomly assigning teachers to classrooms and testing whether the differences in teaching effectiveness estimated when students were assigned "in the usual way" are replicated. In summer 2010, between the first and second year of data collection, roughly 1,600 teachers within each school, grade, and subject essentially drew straws to see which roster of students they would work with during the 2010–11 school year. In our final report, we will describe the degree to which student achievement outcomes following random assignment were consistent with those predicted by the measures of each teacher's effectiveness collected during the prior year, when classrooms were assigned the usual way.

In our final report, we will incorporate data from the National Board for Professional Teaching Standards on applicants from each of the MET project districts; we will add in the results for 9th-grade students; and we will incorporate data from an assessment of teachers' pedagogical content knowledge in math and ELA. In addition, we hope to provide much more specific guidance on the number and duration of observations by school-based personnel required for achieving high levels of reliability.

Referring to the early 2013 results as "final" is a bit of a misnomer. The MET project will be making its data available for other researchers to analyze, which promises years of additional findings. The Inter-University Consortium for Political and Social Research housed at the University of Michigan is creating an archive for storing the data. We hope researchers will replicate the findings above as well as study the many questions we have been unable to address.

ing unconventional teaching methods that don't correspond to the classroom-observation instrument being used in a state or district. If the teacher is getting exemplary student-achievement gains and student survey reports, a school leader should give the teacher the leeway to use a different instructional style. Likewise, if a teacher is performing well on the classroom observations and student surveys but had lower-than-expected student-achievement gains, a school leader might give the teacher the benefit of the doubt for another year and hope that student achievement gains will rise.

Implication for Practice

The MET findings have a number of implications for ongoing efforts to provide more meaningful feedback to teachers:

The main reason to conduct classroom observations is to generate actionable feedback for improving practice. Therefore, the standards need to be clear and the observers should not only be trained, they should demonstrate their understanding of the standards by replicating the ratings given by master scorers. School systems could certify raters using prescored lesson videos, such as we did in our project. They should also conduct multiple observations by more than one rater, and audit a subset of observations to track reliability.

Student surveys are an inexpensive way to add predictive power and reliability to evaluation systems. They could be particularly useful to supplement classroom observations in the grades and subjects where student achievement gains are not available. Although our results suggested such measures could be reliable and predictive, even with students as young as 4th grade, more work needs to be done to evaluate their usefulness in younger grades. To reduce the risk of pressure from teachers or peer pressure from fellow students, it is

important that schools take steps to ensure the anonymity of individual student responses.

When it comes to measuring teachers' effectiveness, the state ELA assessments are less reliable and less related to other measures of practice than state math assessments (or the assessment of students' short-answer writing responses we used to supplement the state tests). The implementation of new literacy assessments in line with the Common Core state standards may help. In the interim, schools might adapt their classroom observations and student surveys to look for evidence of student writing or add questions to the student survey asking students to describe the quality of feedback they receive on their writing.

None of the data collected for MET were used for high-stakes personnel decisions. It may be that the measurement properties of student surveys, or classroom observations, or achievement gains could be distorted when stakes are attached. If principals inflate (or lower) their scores, or if students use the student surveys to play pranks, such changes should become evident in changing relationships among and between the measures. As a result, school systems should monitor those relationships as such systems are implemented.

Finally, we need many more studies evaluating the ways in which better feedback can be paired with targeted development investments to raise teachers' effectiveness in improving student outcomes.

No information is perfect. But better information on teaching effectiveness should allow for improved personnel decisions and faster professional growth. We need to keep in mind the rudimentary indicators used for high-stakes decisions today: teaching experience and educational attainment. When compared with such crude indicators, the combination of student achievement gains on state tests, student surveys, and classroom observations identified teachers with better outcomes on every measure we tested: state tests and supplemental tests as well as more subjective measures, such as student-reported effort and enjoyment in class.

Thomas Kane is professor of education and economics at the Harvard Graduate School of Education. He was formerly deputy director within the U.S. education group at the Bill & Melinda Gates Foundation, where he led the Measures of Effective Teaching project. This essay draws from research done jointly with Douglas O. Staiger from Dartmouth College.