MET project

POLICY AND PRACTICE BRIEF

Ensuring Fair and Reliable Measures of Effective Teaching

Culminating Findings from the MET Project's Three-Year Study

BILL & MELINDA GATES *foundation*

ABOUT THIS REPORT: This non-technical research brief for policymakers and practitioners summarizes recent analyses from the Measures of Effective Teaching (MET) project on identifying effective teaching while accounting for differences among teachers' students, on combining measures into composites, and on assuring reliable classroom observations.[1]

Readers who wish to explore the technical aspects of these analyses may go to www.metproject.org to find the three companion research reports: *Have We Identified Effective Teachers?* by Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger; *A Composite Estimator of Effective Teaching* by Kata Mihaly, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood; and *The Reliability of Classroom Observations by School Personnel* by Andrew D. Ho and Thomas J. Kane.

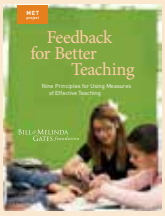
Earlier MET project briefs and research reports also on the website include:

Working with Teachers to Develop Fair and Reliable Measures of Teaching (2010). A white paper describing the rationale for and components of the MET project's study of multiple measures of effective teaching.

Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project (2010). A research report and non-technical policy brief with the same title on analysis of student-perception surveys and student achievement gain measures.

Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains (2012). A research report and policy/practitioner brief with the same title with initial findings on the reliability of classroom observations and implications for combining measures of teaching.

Asking Students about Teaching: Student Perception Surveys and Their Implementation (2012). A non-technical brief for policymakers and practitioners on the qualities of well-designed student surveys and implications for their implementation for teacher feedback and evaluation.



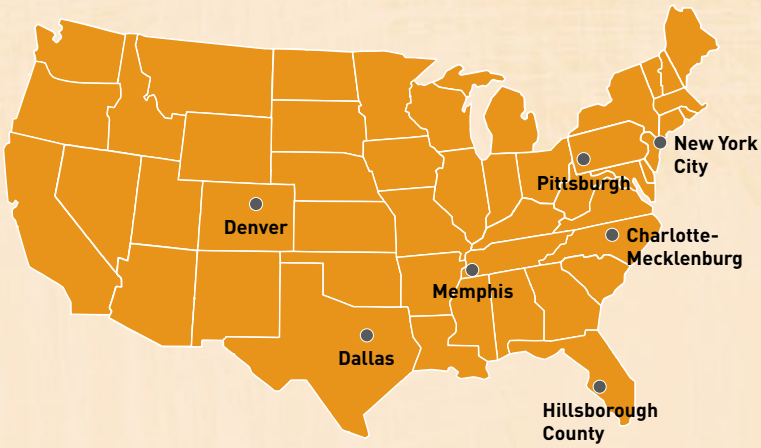
In addition, on www.metproject.org readers will find a set of principles to guide the design of teacher evaluation and support systems based on the work of the MET project, its partners, and other leading systems and organizations, *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching* (2013).

ABOUT THE MET PROJECT: The MET project is a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Funding is provided by the Bill & Melinda Gates Foundation.

The approximately 3,000 MET project teachers who volunteered to open up their classrooms for this work are from the following districts: The Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County Public Schools, the Memphis Public Schools, the New York City Schools, and the Pittsburgh Public Schools.

Partners include representatives of the following institutions and organizations: American Institutes for Research, Cambridge Education, University of Chicago, The Danielson Group, Dartmouth College, Educational Testing Service, Empirical Education, Harvard University, National Board for Professional Teaching Standards, National Math and Science Initiative, New Teacher Center, University of Michigan, RAND, Rutgers University, University of Southern California, Stanford University, Teachscape, University of Texas, University of Virginia, University of Washington, and Westat.

MET Project Teachers



January 2013

Contents

Executive Summary ____ 3

Can Measures of Effective Teaching Identify Teachers Who Better Help Students Learn? ____ 6

How Much Weight Should Be Placed on Each Measure of Effective Teaching? ____ 10

How Can Teachers Be Assured Trustworthy Results from Classroom Observations? ____ 16

What We Know Now ____ 20

Endnotes ____ 23

Executive Summary

States and districts have launched unprecedented efforts in recent years to build new feedback and evaluation systems that support teacher growth and development. The goal is to improve practice so that teachers can better help their students graduate from high school ready to succeed in college and beyond.

These systems depend on trustworthy information about teaching effectiveness—information that recognizes the complexity of teaching and is trusted by both teachers and administrators. To that end, the Measures of Effective Teaching (MET) project set out three years ago to investigate how a set of measures could identify effective teaching fairly and reliably. With the help of 3,000 teacher volunteers who opened up their classrooms to us—along with scores of academic and organizational partners—we have studied, among other measures:

- **Classroom observation instruments,** including both subject-specific and cross-subject tools, that define discrete teaching competencies and describe different levels of performance for each;
- **Student perception surveys** that assess key characteristics of the classroom environment, including supportiveness, challenge, and order; and
- **Student achievement gains** on state tests and on more cognitively challenging assessments.

We have reported findings as we learned them in order to provide states and districts with evidence-based guidance to inform their ongoing work. In our initial report in 2010 (*Learning about Teaching*), we found that a well-designed student perception survey can provide reliable feedback on aspects of teaching practice that are predictive of student learning. In 2012 (*Gathering Feedback for Teaching*), we presented similar results for classroom observations. We also found that an accurate observation rating requires two or more lessons, each scored by a different certified observer. With each analysis we have better understood the particular contribution that each measure makes to a complete picture of effective teaching and how those measures should be implemented to provide teachers with accurate and meaningful feedback.

This final brief from the MET project's three-year study highlights new analyses that extend and deepen the insights from our previous work. These studies address three fundamental questions that face practitioners and policymakers engaged in creating teacher support and evaluation systems.

"Feedback and evaluation systems depend on trustworthy information about teaching effectiveness to support improvement in teachers' practice and better outcomes for students."

The Questions

Can measures of effective teaching identify teachers who better help students learn?

Despite decades of research suggesting that teachers are the most important in-school factor affecting student learning, an underlying question remains unanswered: Are seemingly more effective teachers truly better than other teachers at improving student learning, or do they simply have better students?

Ultimately, the only way to resolve that question was by randomly assigning students to teachers to see if teachers previously identified as more effective actually caused those students to learn more. That is what we did for a subset of MET project teachers. Based on data we collected during the 2009–10 school year, we produced estimates of teaching effectiveness for each teacher. We adjusted our estimates to account for student differences in prior test scores, demographics, and other traits. We then randomly assigned a classroom of students to each participating teacher for 2010–11.

Following the 2010–11 school year we asked two questions: First, did students actually learn more when randomly assigned to the teachers who seemed more effective when we evaluated them the prior year? And, second, did the magnitude of the difference in student outcomes following random assignment correspond with expectations?

How much weight should be placed on each measure of effective teaching?

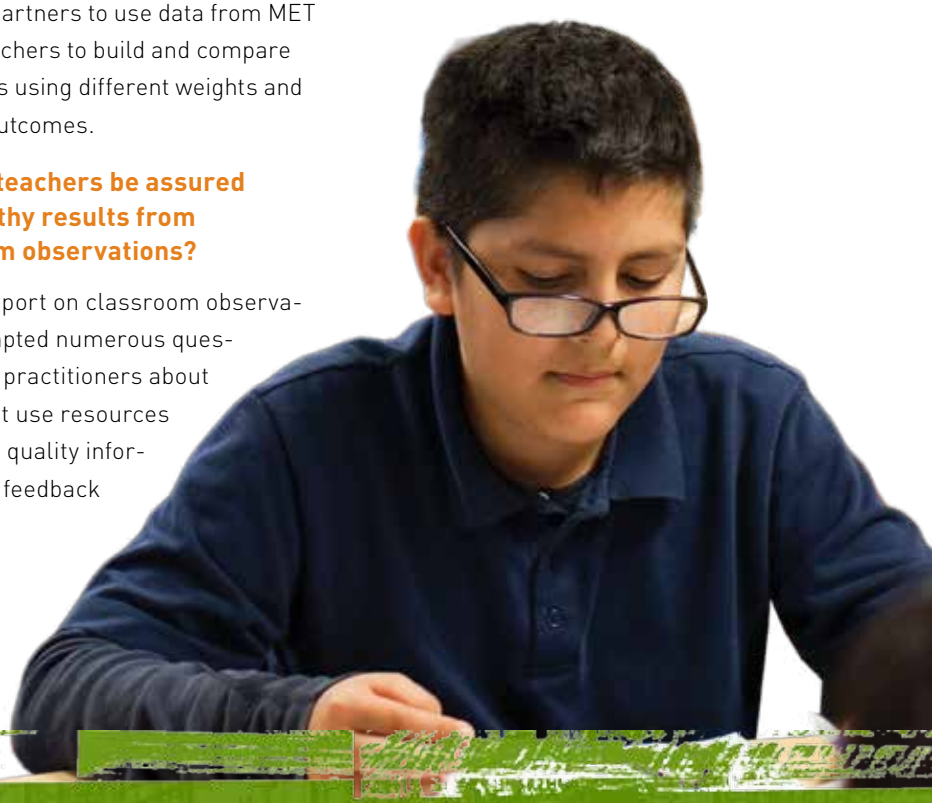
While using multiple measures to provide feedback to teachers, many states and districts also are combining measures into a single index to support decisionmaking. To date, there has been little empirical evidence to inform how systems might weight each measure within a composite to support improvements in teacher effectiveness. To help fill that void, we tasked a group of our research partners to use data from MET project teachers to build and compare composites using different weights and different outcomes.

How can teachers be assured trustworthy results from classroom observations?

Our last report on classroom observations prompted numerous questions from practitioners about how to best use resources to produce quality information for feedback on classroom practice. For example: How many observers are needed to achieve sufficient reliability from a given number of observations? Do all observations need to be the same length to have confidence in the results? And what is the value of adding observers from outside a teacher's own school? To help answer these questions, we designed a study in which administrators and peer observers produced more than 3,000 scores for lessons taught by teachers within one MET project partner school district.

Key findings from those analyses:

1. **Effective teaching can be measured.** We collected measures of teaching during 2009–10. We adjusted those measures for the backgrounds and prior achievement of the students in each class. But, without random assignment, we had no way to know if the adjustments we made were sufficient to discern the markers of effective teaching from the unmeasured aspects of students' backgrounds.



In fact, we learned that the adjusted measures did identify teachers who produced higher (and lower) average student achievement gains following random assignment in 2010–11. The data show that we can identify groups of teachers who are more effective in helping students learn. Moreover, the magnitude of the achievement gains that teachers generated was consistent with expectations.

In addition, we found that more effective teachers not only caused students to perform better on state tests, but they also caused students to score higher on other, more cognitively challenging assessments in math and English.

2. **Balanced weights indicate multiple aspects of effective teaching.** A composite with weights between 33 percent and 50 percent assigned to state test scores demonstrated the best mix of low volatility from year to year and ability to predict student gains on multiple assessments. The composite that best indicated improvement on state tests heavily weighted teachers' prior student achievement gains based on those same tests. But composites that assigned 33 percent to 50 percent of the weight to state tests did nearly as well and were somewhat better at predicting student learning on more cognitively challenging assessments.

   Multiple measures also produce more consistent ratings than student achievement measures alone. Estimates of teachers' effectiveness are more stable from year to year when they combine classroom observations, student surveys, and measures of student achievement gains than when they are based solely on the latter.

3. **Adding a second observer increases reliability significantly more than having the same observer score an additional lesson.** Teachers' observation scores vary more from observer to observer than from lesson to lesson. Given the same total number of observations, including the perspectives of two or more observers per teacher greatly enhances reliability. Our study of video-based observation scoring also revealed that:

   a. Additional shorter observations can increase reliability. Our analysis suggests that having additional observers watch just part of a lesson may be a cost-effective way to boost reliability by including additional perspectives.

   b. Although school administrators rate their own teachers somewhat higher than do outside observers, how they rank their teachers' practice is very similar and teachers' own administrators actually discern bigger differences in teaching practice, which increases reliability.

   c. Adding observations by observers from outside a teacher's school to those carried out by a teacher's own administrator can provide an ongoing check against in-school bias. This could be done for a sample of teachers rather than all, as we said in *Gathering Feedback for Teaching*.

The following pages further explain these findings and the analyses that produced them.

Can Measures of Effective Teaching Identify Teachers Who Better Help Students Learn?[2]

By definition, teaching is effective when it enables student learning. But identifying effective teaching is complicated by the fact that teachers often have very different students. Students start the year with different achievement levels and different needs. Moreover, some teachers tend to get particular types of students year after year (that is, they tend to get higher-performing or lower-performing ones). This is why so-called value-added measures attempt to account for differences in the measurable characteristics of a teacher's students, such as prior test scores and poverty.

"Teachers previously identified as more effective caused students to learn more. Groups of teachers who had been identified as less effective caused students to learn less."

However, students differ in other ways—such as behavior and parental involvement—which we typically cannot account for in determining teaching effectiveness. If those "unaccounted for" differences also affect student learning, then what seems like effective teaching may actually reflect unmeasured characteristics of a teacher's students. The only way to know if measures of teaching truly identify effective teaching and not some unmeasured student characteristics is by randomly assigning teachers to students. So we did.

In 2009–10, we measured teachers' effectiveness using a combined measure, comprising teachers' classroom observation results, student perception survey responses, and student achievement gains adjusted for student characteristics, such as prior performance and demographics. The following year (2010–11), we randomly assigned different rosters of students to two or more MET project teachers who taught the same grade and subject in the same school. Principals created rosters and the RAND Corp assigned them randomly to teachers (see **Figure 1**). Our aim was to determine if the students who were randomly assigned to teachers who previously had been identified as more effective actually performed better at the end of the 2010–11 school year.[3]

They did. On average, the 2009–10 composite measure of effective teaching accurately predicted 2010–11 student performance. The research confirmed that, as a group, teachers previously identified as more effective caused students to learn more. Groups of teachers who had been identified as less effective

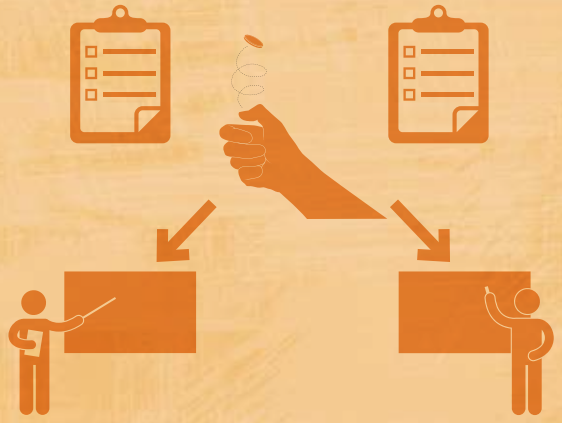
caused students to learn less. We can say they "caused" more (or less) student learning because when we randomly assigned teachers to students during the second year, we could be confident that any subsequent differences in achievement were being driven by the teachers, not by the unmeasured characteristics of their students. In addition, the magnitude of the gains they caused was consistent with our expectations.

Figure 2 illustrates just how well the measures of effective teaching predicted student achievement following random assignment. The diagonal line represents perfect prediction. Dots above the diagonal line indicate groups of teachers whose student outcomes following random assignment were better than predicted. Dots below the line indicate groups of teachers whose student outcomes following random assignment were worse than predicted. Each dot

Figure 1

Putting Measures of Effective Teaching to the Test with Random Assignment

1. Principals created rosters for each class
2. The rosters were assigned randomly within each grade and subject
3. We predicted student outcomes based on teachers' previous results, observations, and student surveys.
4. We compared those predictions to actual differences.



Do measures of teaching really identify teachers who help students learn more, or do seemingly more effective teachers just get better students? To find out, the MET project orchestrated a large-scale experiment with MET project teachers to see if teachers identified as more effective than their peers would have greater student achievement gains even with students who were assigned randomly.

To do so, the MET project first estimated teachers' effectiveness using multiple measures from the 2009–10 school year. As is common in schools, some teachers had been assigned students with stronger prior achievement than others. In assessing each teacher's practice that year, the project controlled for students' prior achievement and demographic characteristics. But there may have been other differences among students as well. So for the following school year (2010–11), principals created rosters of students for each class in the study, and then researchers randomly assigned each roster to a participating teacher from among those who could teach the class.

At the end of the 2010–11 school year, MET project analysts checked to see if students taught by teachers identified as more effective than their colleagues actually had greater achievement gains than students taught by teachers identified as less effective. They also checked to see how well actual student achievement gains for teachers matched predicted gains.

represents 5 percent of the teachers in the analysis, sorted based on their predicted impact on student achievement.[4]

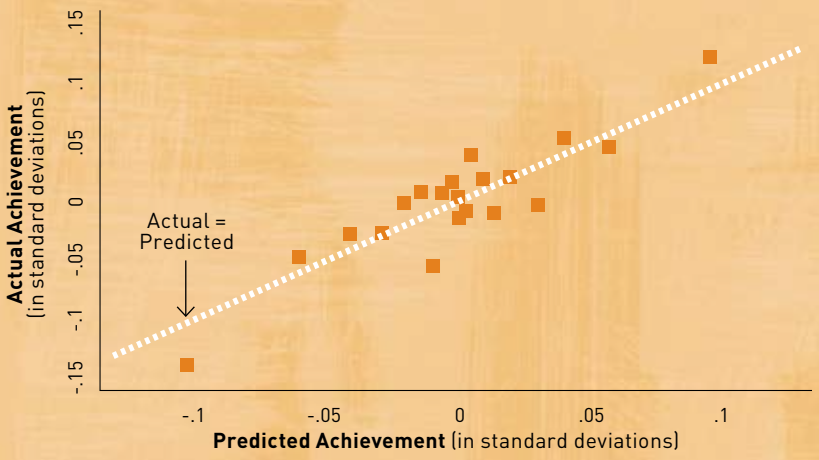
As seen in **Figure 2**, in both math and English language arts (ELA), the groups of teachers with greater predicted impacts on student achievement generally had greater actual impacts on student achievement following random assignment. Further, the actual impacts are approximately in line with the predicted impacts.[5] We also found that teachers who we identified as being effective in promoting achievement on the state tests also generated larger gains on the supplemental tests administered in spring 2011.

Based on our analysis, we can unambiguously say that school systems should account for the prior test scores of students. When we removed this control, we wound up predicting much larger differences in achievement than actually occurred, indicating that student assignment biased the results. However, our analysis could not shed as much light on the need to control for demographics or "peer effects"—that is, the average prior achievement and demographics of each student's classmates. Although we included those

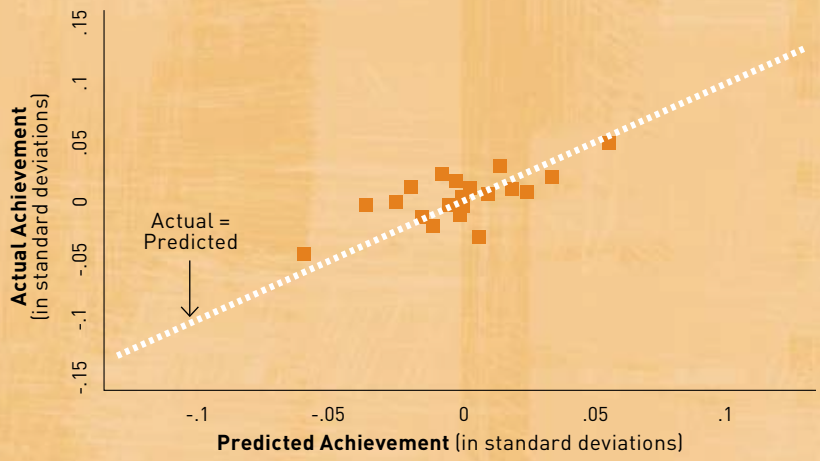
Figure 2

Effectiveness Measures Identify Teachers Who Help Students Learn More

Actual and Predicted Achievement of Randomized Classrooms (Math)



Actual and Predicted Achievement of Randomized Classrooms (English Language Arts)



These charts compare the actual 2010–11 school year achievement gains for randomly assigned classrooms with the results that were predicted based on the earlier measures of teaching effectiveness. Each dot represents the combination of actual and estimated student performance for 5 percent of the teachers in the study, grouped by the teachers' estimated effectiveness. The dashed line shows where the dots would be if the actual and predicted gains matched perfectly.

On average, students of teachers with higher teacher effectiveness estimates outperformed students of teachers with lower teacher effectiveness estimates. Moreover, the magnitude of students' actual gains largely corresponded with gains predicted by their effectiveness measured the previous year. Both the actual and predicted achievement are reported relative to the mean in the randomization block. That is, a zero on either axis implies that the value was no different from the mean for the small group of teachers in a grade, subject, and school within which class lists were randomized.

Impacts are reported in student-level standard deviations. A .25 standard deviation difference is roughly equivalent to a year of schooling. The predicted impacts are adjusted downward to account for incomplete compliance with randomization.

"We can unambiguously say that school systems should adjust their achievement gain measures to account for the prior test scores of students. When we removed this control, we wound up predicting much larger differences in achievement than actually occurred."

controls, we cannot determine from our evidence whether school systems should include them. Our results were ambiguous on that score.

To avoid over-interpretation of these results, we hasten to add two caveats: First, a prediction can be correct on average but still be subject to measurement error. Our predictions of students' achievement following random assignment were correct on average, but within every group there were some teachers whose students performed better than predicted and some whose students performed worse. Second, we could not, as a practical matter, randomly assign students or teachers to a different school site. As a result, our study does not allow us to investigate bias in teacher effectiveness measures arising from student sorting between different schools.[6]

Nonetheless, our analysis should give heart to those who have invested considerable effort to develop practices and policies to measure and support effective teaching. Through this large-scale study involving random assignment of teachers to students, we are confident that we can identify groups of teachers who are comparatively more effective than their peers in helping students learn. Great teaching does make a difference.

How Much Weight Should Be Placed on Each Measure of Effective Teaching?[7]

Teaching is too complex for any single measure of performance to capture it accurately. Identifying great teachers requires multiple measures. While states and districts embrace multiple measures for targeted feedback, many also are combining measures into a single index, or composite. An index or composite can be a useful summary of complex information to support decisionmaking. The challenge is to combine measures in ways that support effective teaching while avoiding such unintended consequences as too-narrow a focus on one aspect of effective teaching.

To date, there has been little empirical evidence to suggest a rationale for particular weights. The MET project's report *Gathering Feedback for Teaching* showed that equally weighting three measures, including achievement gains, did a better job predicting teachers' success (across several student outcomes) than teachers' years of experience and masters' degrees. But that work did not attempt to determine optimal weights for composite measures.

Over the past year, a team of MET project researchers from the RAND Corporation and Dartmouth College used MET project data to compare differently weighted composites and study the implications of different weighting schemes for different outcomes. As in the *Gathering Feedback for Teaching* report, these composites included student achievement gains based on state assessments, classroom observations, and student surveys. The researchers estimated the ability of variously weighted composites to produce consistent results and accurately forecast teachers' impact on student achievement gains on different types of tests.

The goal was not to suggest a specific set of weights but to illustrate the trade-offs involved when choosing weights. Assigning significant weight to one measure might yield the best predictor of future performance on that measure. But heavily weighting a single measure may incentivize teachers to focus too narrowly on a single aspect

of effective teaching and neglect its other important aspects. For example, a singular focus on state tests could displace gains on other harder-to-measure outcomes. Moreover, if the goal is for students to meet a broader set of learning objectives than are measured by a state's tests, then too-heavily weighting that test could make it harder to identify teachers who are producing other valued outcomes.

Composites Compared

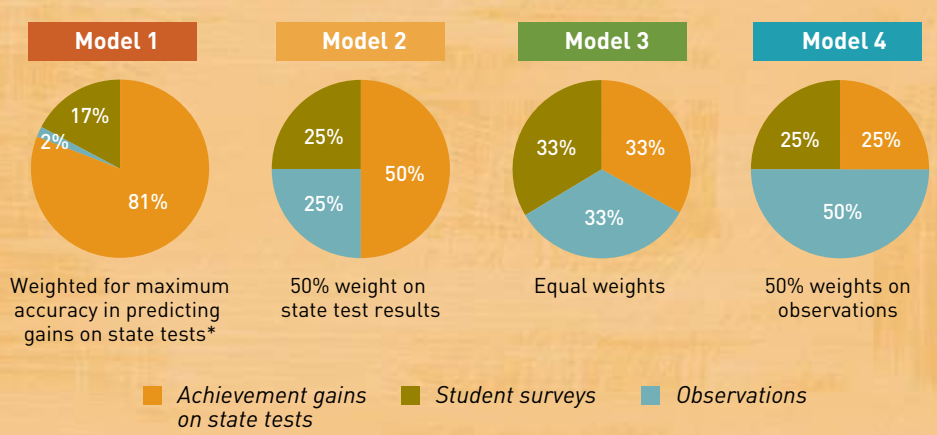
The research team compared four different weighting models, illustrated in **Figure 3:** (Model 1) The "best predictor" of state achievement test gains (with weights calculated to maximize the ability to predict teachers' student achievement gains on state tests, resulting in 65+ percent of the weight being placed on the student achievement gains across grades and subjects); (Model 2) a composite that assigned 50 percent of the weight to students' state achievement test gains; (Model 3) a composite that applied equal weights to each measure; and (Model 4) one that gave 50 percent to observation ratings and 25 percent each to achievement gains and student surveys. The weights that best predict state tests, shown for Model 1 in **Figure 3,** were calculated to predict gains on state ELA tests at the middle school level, which assigns a whopping 81 percent of the weight to prior gains on the same tests (best-predictor weights for other grades and subjects are in the table on page 14).

Figure 4 compares the different weighting schemes on three criteria, using middle school ELA as an example (see the table on page 14 for other grades and subjects). The first is predicting teachers' student achievement gains on state assessments. A correlation of 1.0 would indicate perfect accuracy in

> "Heavily weighting a single measure may incentivize teachers to focus too narrowly on a single aspect of effective teaching and neglect its other important aspects. ... [I]f the goal is for students to meet a broader set of learning objectives than are measured by a state's tests, then too-heavily weighting that test could make it harder to identify teachers who are producing other valued outcomes."

Figure 3

Four Ways to Weight



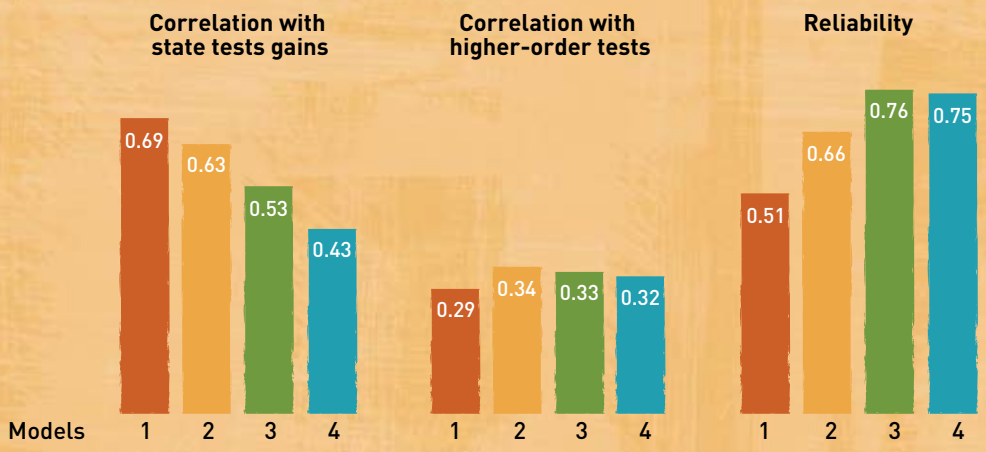
**Weights shown for Model 1 were calculated to best predict gains on state tests for middle school English language arts. Similar best predictor weights for other grades and subjects are in the table on page 14.*

These charts illustrate four ways to construct a composite measure of effective teaching. Each model uses different weights but includes the same components— student achievement gains on the state tests, student perception surveys, and classroom observations. Model 1 uses the weights that would best predict a teacher's impact on state test scores. Across grades and subjects, the "best predictor" model assigns 65 percent or more of the weight to a teacher's prior state test gains. Models 2–4 are not based on maximizing any particular outcome. They approximate different weighting schemes used by states and districts, with each model placing progressively less weight on student achievement gains on state tests.

Figure 4

Trade-Offs from Different Weighting Schemes

Middle School English Language Arts



These bars compare the four weighting schemes in Figure 3 on three criteria: accuracy in predicting teachers' achievement gains on state tests; accuracy in predicting student achievement gains on supplemental assessments designed to test higher-order thinking skills; and reliability, reflecting the year-to-year stability of teachers' results. Shown are the results for middle school ELA (see Table 1 on page 14 for results for other grades and subjects).

As indicated, Model 2 (50 percent state test results) and Model 3 (33 percent state tests) achieve much of the same predictive power as Model 1 (the "best predictor" of state test results) in anticipating teachers' future state test results (Model 1). Model 4 (50 percent observation) is considerably less predictive. However, the figures also illustrate two other trade-offs. Models 2 and 3 also are somewhat better than Model 1 at predicting gains on the tests of higher-order thinking skills (for all but elementary school math). Across most grades and subjects, Model 1 was the least reliable.

predicting teachers' student achievement gains on state tests. By definition, the best composite in this regard is Model 1, the model weighted for maximizing accuracy on state test results. Models 2–4 show the effect of reducing weights on student achievement gains on state tests for middle school ELA. As shown from middle school ELA, reducing weights on student achievement gains decreases the power to predict future student achievement gains on state tests from 0.69 to 0.63 with Model 2; to 0.53 with Model 3; and to 0.43 with Model 4. Other grades and subjects showed similar patterns, as indicated in the table on page 14.

While it is true that the state tests are limited and that schools should value other outcomes, observations and student surveys may not be more correlated with those other outcomes than the state tests. As a result, we set out to test the strength of each model's correlation with another set of test outcomes. The middle set of bars in **Figure 4** compares the four models (see Figure 3)—each using state test results to measure achievement gains—on how well they would predict teachers' student achievement gains on supplemental tests that were administered in MET project teachers' classrooms: The SAT 9 Open-Ended Reading Assessment (SAT 9 OE) and the Balanced Assessment in Mathematics (BAM).

While covering less material than state tests, the SAT 9 OE and BAM assessments include more cognitively challenging items that require writing, analysis, and application of concepts, and they are meant to assess higher-order thinking skills. Sample items released by the assessment consortia for the new Common Core State Standards assessments are more similar to the items on these supplemental tests than the ones on the state assessments. Shown in **Figure 4** is the effect of reducing the weight on state test gains in predicting gains on these other assessments, again for middle school ELA. For most grades and subjects, Model 2 and Model 3 (50 percent state test and equal weights for all three measures) best predicted teachers' student achievement gains on these supplemental assessments, with little difference between the two models. The one exception was elementary school math, where Model 1 (best predictor) was best.

The third set of bars in **Figure 4** compares composites on their reliability—that is, the extent to which the composite would produce consistent results for the same teachers from year to year (on a scale from 0–1.0, with

Increasing Accuracy, Reducing Mistakes

When high-stakes decisions must be made, can these measures support them? Undoubtedly, that question will be repeated in school board meetings and in faculty break rooms around the country in the coming years.

The answer is yes, not because the measures are perfect (they are not), but because the combined measure is better on virtually every dimension than the measures in use now. There is no way to avoid the stakes attached to every hiring, retention, and pay decision. And deciding not to make a change is, after all, a decision. No measure is perfect, but better information should support better decisions.

In our report *Gathering Feedback for Teaching,* we compared the equally weighted measure (Model 3 in Figures 3 and 4) to two indicators that are almost universally used for pay or retention decisions today: teaching experience and possession of a master's degree. On every student outcome—the state tests, supplemental tests, student's self-reported level of effort and enjoyment in class—the teachers who excelled on the composite measure had better outcomes than those with high levels of teaching experience or a master's degree.

In addition, many districts currently require classroom observations, but they do not include student surveys or achievement gains. We tested whether observations alone are enough. Even with four full classroom observations (two by one observer and two by another), conducted by observers trained and certified by the Educational Testing Service, the observation-only model performed far worse than any of our multiple measures composites. (The correlations comparable to those in Figure 5 would have been .14 and .25 with the state tests and test of higher-order skills.)

Still, it is fair to ask, what might be done to reduce error? Many steps have been discussed in this and other reports from the project:

- First, if any type of student data is to be used—either from tests or from student surveys—school systems should give teachers a chance to correct errors in their student rosters.
- Second, classroom observers should not only be trained on the instrument. They should first demonstrate their accuracy by scoring videos or observing a class with a master observer.
- Third, observations should be done by more than one observer. A principal's observation is not enough. To ensure reliability, it is important to involve at least one other observer, either from inside or outside the school.
- Fourth, if multiple years of data on student achievement gains, observations, and student surveys are available, they should be used. For novice teachers and for systems implementing teacher evaluations for the first time, there may be only a single year available. We have demonstrated that a single year contains information worth acting on. But the information would be even better if it included multiple years. When multiple years of data are available they should be averaged (although some systems may choose to weight recent years more heavily).

1.0 representing perfect consistency and no volatility). Again, results shown are for middle school ELA. Across all grades and subjects, the most reliable composites were either Models 2 (50 percent state test) or 3 (equal weights). For all but middle school math, the least reliable composite was Model 1 (best predictor). Model 4 (50 percent observations) was somewhat less reliable than Model 2 (equal weights) for all grades and subjects. Although not shown, student achievement gains on state tests by themselves are less stable than all of the composites, with one exception: Model 4 (50 percent observations) is slightly less stable than achievement gains alone for middle school math.

General Implications

The intent of this analysis was not to recommend an ideal set of weights to use in every circumstance. Rather, our goal was to describe the trade-offs among different approaches.[8]

If the goal is to predict gains on state tests, then the composites that put 65+ percent of the weight on the student achievement gains on those tests will generally show the greatest accuracy. However, reducing the weights on the state test achievement gain measures to 50 percent or 33 percent generates two positive trade-offs: it increases stability (lessens volatility from year to year) and it also increases somewhat the correlation with tests other than the state tests.

However, it is possible to go too far. Lowering the weight on state test achievement gains below 33 percent, and raising the weight on observations to 50 percent and including student surveys at 25 percent, is counterproductive. It not only lowers the

Table 1

CALCULATED WEIGHTS FOR MAXIMUM ACCURACY IN PREDICTING GAINS ON STATE TESTS

	English Language Arts			Math		
	State Tests	Observations	Student Surveys	State Tests	Observations	Student Surveys
Elementary	65%	9%	25%	85%	5%	11%
Middle	81%	2%	17%	91%	4%	5%

RELIABILITY AND ACCURACY OF DIFFERENT WEIGHTING SCHEMES

		English Language Arts				Math			
		Weighted for Max State Test Accuracy	50% State Test	Equal Weights	50% Observations	Weighted for Max State Test Accuracy	50% State Test	Equal Weights	50% Observations
Elementary	Reliability	0.42	0.46	0.50	0.49	0.52	0.57	0.57	0.55
	Correlation with state test	0.61	0.59	0.53	0.45	0.72	0.65	0.54	0.46
	Correlation with higher-order test	0.35	0.37	0.37	0.35	0.31	0.29	0.25	0.20
Middle	Reliability	0.51	0.66	0.76	0.75	0.86	0.88	0.88	0.83
	Correlation with state test	0.69	0.63	0.53	0.43	0.92	0.84	0.73	0.65
	Correlation with higher-order test	0.29	0.34	0.33	0.32	0.38	0.44	0.45	0.45

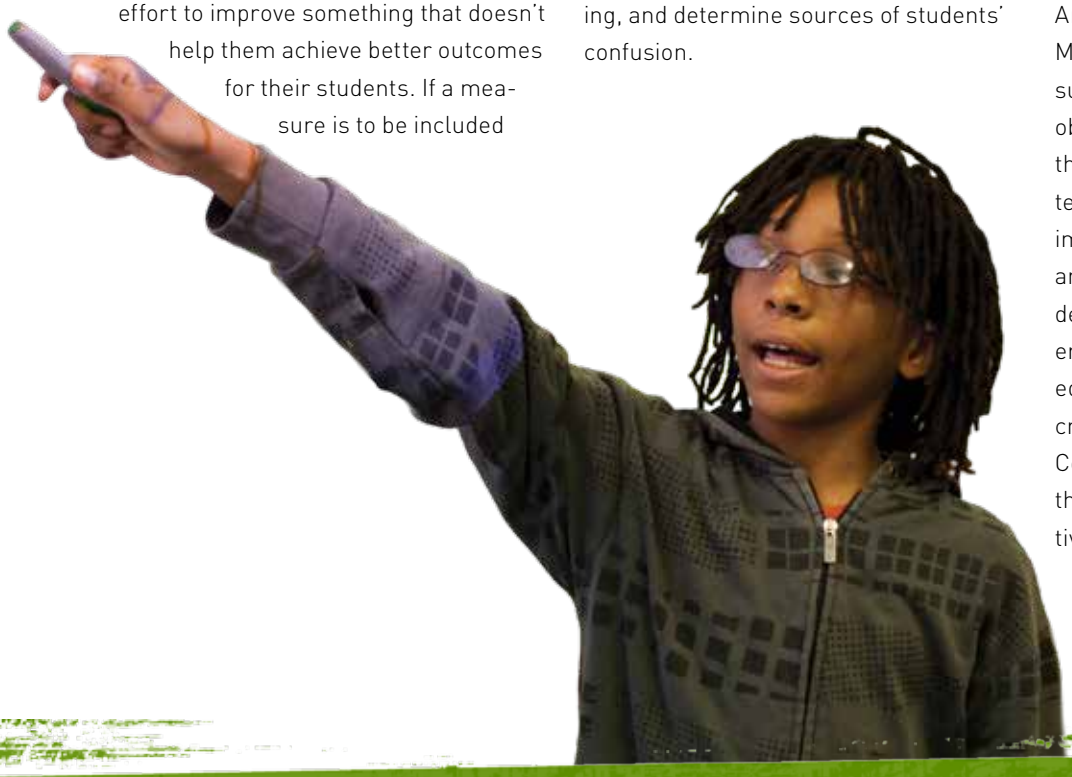
correlation with state achievement gains; it can also lower reliability and the correlation with other types of testing outcomes.

Ultimately, states, local education authorities, and other stakeholders need to decide how to weight the measures in a composite. Our data suggest that assigning 50 percent or 33 percent of the weight to state test results maintains considerable predictive power, increases reliability, and potentially avoids the unintended negative consequences from assigning too-heavy weights to a single measure. Removing too much weight from state tests, however, may not be a good idea, given the lower predictive power and reliability of Model 4 (25 percent state tests). In short, there is a range of reasonable weights for a composite of multiple measures.

Validity and Content Knowledge for Teaching

Teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes. This test for "validity" has been central to the MET project's analyses. Measures that have passed this test include high-quality classroom observations, well-designed student-perception surveys, and teachers' prior records of student achievement gains on state tests.

Over the past year, MET project researchers have investigated another type of measure, called the Content Knowledge for Teaching (CKT) tests. These are meant to assess teachers' understanding of how students acquire and understand subject-specific skills and concepts in math and ELA. Developed by the Educational Testing Service and researchers at the University of Michigan, these tests are among the newest measures of teaching included in the MET project's analyses. Mostly multiple choice, the questions ask how to best represent ideas to students, assess student understanding, and determine sources of students' confusion.



The CKT tests studied by the MET project did not pass our test for validity. MET project teachers who performed better on the CKT tests were not substantively more effective in improving student achievement on the outcomes we measured. This was true whether student achievement was measured using state tests or the supplemental assessments of higher-order thinking skills. For this reason, the MET project did not include CKT results within its composite measure of effective teaching.

These results, however, speak to the validity of the current measure still early in its development in predicting achievement gains on particular student assessments—not to the importance of content-specific pedagogical knowledge. CKT as a concept remains promising. The teachers with higher CKT scores did seem to have somewhat higher scores on two subject-based classroom observation instruments: the Mathematical Quality of Instruction (MQI) and the Protocol for Language Arts Teacher Observations (PLATO). Moreover, the MET project's last report suggested that some content-specific observation instruments were better than cross-subject ones in identifying teachers who were more effective in improving student achievement in ELA and math. Researchers will continue to develop measures for assessing teachers' content-specific teaching knowledge and validating them as states create new assessments aligned to the Common Core State Standards. When they have been shown to be substantively related to a teacher's students' achievement gains, these should be considered for inclusion as part of a composite measure of effective teaching.

How Can Teachers Be Assured Trustworthy Results from Classroom Observations?[9]

Classroom observations can be powerful tools for professional growth. But for observations to be of value, they must reliably reflect what teachers do throughout the year, as opposed to the subjective impressions of a particular observer or some unusual aspect of a particular lesson. Teachers need to know they are being observed by the right people, with the right skills, and a sufficient number of times to produce trustworthy results. Given this, the challenge for school systems is to make the best use of resources to provide teachers with high-quality feedback to improve their practice.

"For the same total number of observations, incorporating additional observers increases reliability."

The MET project's report *Gathering Feedback for Teaching* showed the importance of averaging together multiple observations from multiple observers to boost reliability. Reliability represents the extent to which results reflect consistent aspects of a teacher's practice, as opposed to other factors such as observer judgment. We also stressed that observers must be well-trained and assessed for accuracy before they score teachers' lessons.

But there were many practical questions the MET project couldn't answer in its previous study. Among them:

- Can school administrators reliably assess the practice of teachers in their schools?
- Can additional observations by external observers not familiar with a teacher increase reliability?
- Must all observations involve viewing the entire lesson or can partial lessons be used to increase reliability? And,
- What is the incremental benefit of adding additional lessons and additional observers?

These questions came from our partners, teachers, and administrators in urban school districts. In response, with the help of a partner district, the Hillsborough County (Fla.) Public Schools, the MET project added a study of classroom observation

Hillsborough County's Classroom Observation Instrument

Like many school districts, Hillsborough County uses an evaluation instrument adapted from the Framework for Teaching, developed by Charlotte Danielson. The framework defines four levels of performance for specific competencies in four domains of practice. Two of those domains pertain to activities outside the classroom: Planning and Preparation, and Professional Responsibility. Observers rated teachers on the 10 competencies in the framework's two classroom-focused domains, as shown:

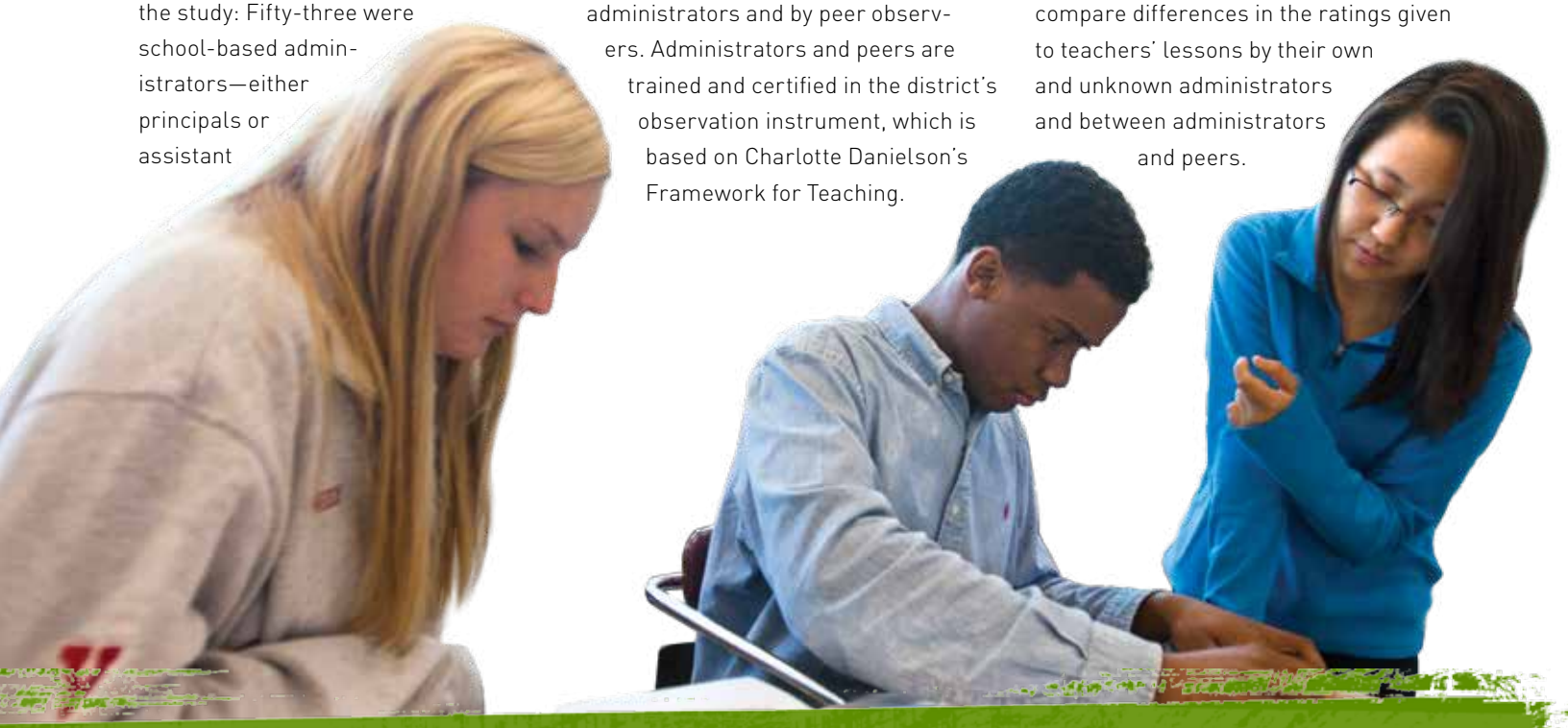
Domain 2: The Classroom Environment	Domain 3: Instruction
▪ Creating an Environment of Respect and Rapport	▪ Communicating with Students
▪ Establishing a Culture of Learning	▪ Using Discussion and Questioning Techniques
▪ Managing Classroom Procedures	▪ Engaging Students in Learning
▪ Managing Student Behavior	▪ Using Assessment in Instruction
▪ Organizing Physical Space	▪ Demonstrating Flexibility and Responsiveness

reliability. This study engaged district administrators and teacher experts to observe video-recorded lessons of 67 Hillsborough County teachers who agreed to participate.

Comparison of Ratings

Two types of observers took part in the study: Fifty-three were school-based administrators—either principals or assistant principals—and 76 were peer observers. The latter are district-based positions filled by teachers on leave from the classroom who are responsible for observing and providing feedback to teachers in multiple schools. In Hillsborough County's evaluation system, teachers are observed multiple times, formally and informally, by their administrators and by peer observers. Administrators and peers are trained and certified in the district's observation instrument, which is based on Charlotte Danielson's Framework for Teaching.

These observers each rated 24 lessons for us and produced more than 3,000 ratings that we could use to investigate our questions. MET project researchers were able to calculate reliability for many combinations of observers (administrator and peer), lessons (from 1 to 4), and observation duration (full lesson or 15 minutes). We were able to compare differences in the ratings given to teachers' lessons by their own and unknown administrators and between administrators and peers.



Effects on Reliability

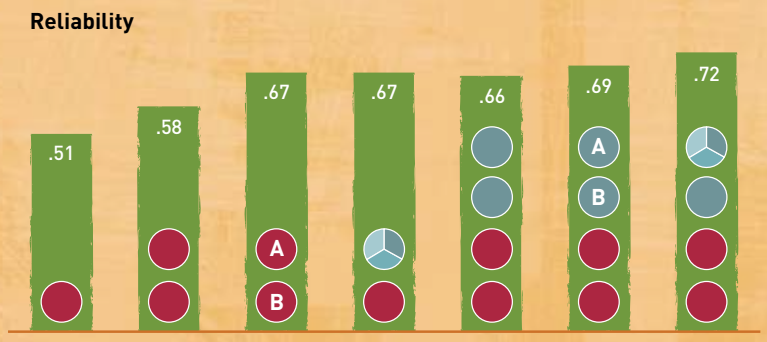
Figure 5 graphically represents many of the key findings from our analyses of those ratings. Shown are the estimated reliabilities for results from a given set of classroom observations. Reliability is expressed on a scale from 0 to 1. A higher number indicates that results are more attributable to the particular teacher as opposed to other factors such as the particular observer or lesson. When results for the same teachers vary from lesson to lesson or from observer to observer, then averaging teachers' ratings across multiple lessons or observers decreases the amount of "error" due to such factors, and it increases reliability.

Adding lessons and observers increases the reliability of classroom observations. In our estimates, if a teacher's results are based on two lessons, having the second lesson scored by a second observer can boost reliability significantly. This is shown in **Figure 5:** When the same administrator observes a second lesson, reliability increases from .51 to .58, but when the second lesson is observed by a different administrator from the same school, reliability increases more than twice as much, from .51 to .67. Whenever a given number of lessons was split between multiple observers, the reliability was greater than that achieved by a single observer. In other words, for the same total number of observations, incorporating additional observers increases reliability.

Of course, it would be a problem if school administrators and peer observers produced vastly different results for the same teachers. But we didn't find that to be the case. Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by peer observers and administrators from other schools. This implies that administrators are seeing the same

Figure 5

There Are Many Roads to Reliability



Lesson observed by own administrator = 45 min
Lesson observed by peer observer = 45 min
Three 15-minute lessons observed by three additional peer observers = 45 min
A and **B** denote different observers of the same type

These bars show how the number of observations and observers affects reliability. Reliability represents the extent to which the variation in results reflects consistent aspects of a teacher's practice, as opposed to other factors such as differing observer judgments. Different colors represent different categories of observers. The "A" and "B" in column three show that ratings were averaged from two different own-school observers. Each circle represents approximately 45 minutes of observation time (a solid circle indicates one observation of that duration, while a circle split into three indicates three 15-minute observations by three observers). As shown, reliabilities of .66–.72 can be achieved in multiple ways, with different combinations of number of observers and observations. (For example, one observation by a teacher's administrator when combined with three short, 15-minute observations each by a different observer would produce a reliability of .67.)

things in the videos that others do, and they are not being swayed by personal biases.

If additional observations by additional observers are important, how can the time for those added observations be divided up to maximize the use of limited resources while assuring trustworthy results? This is an increasingly relevant question as more school systems make use of video in providing teachers with feedback on their practice. Assuming multiple videos for a teacher exist, an observer could use the same amount of time to watch one full lesson or two or three partial lessons. But to consider the latter, one would want to know whether partial-lesson observations increase reliability.

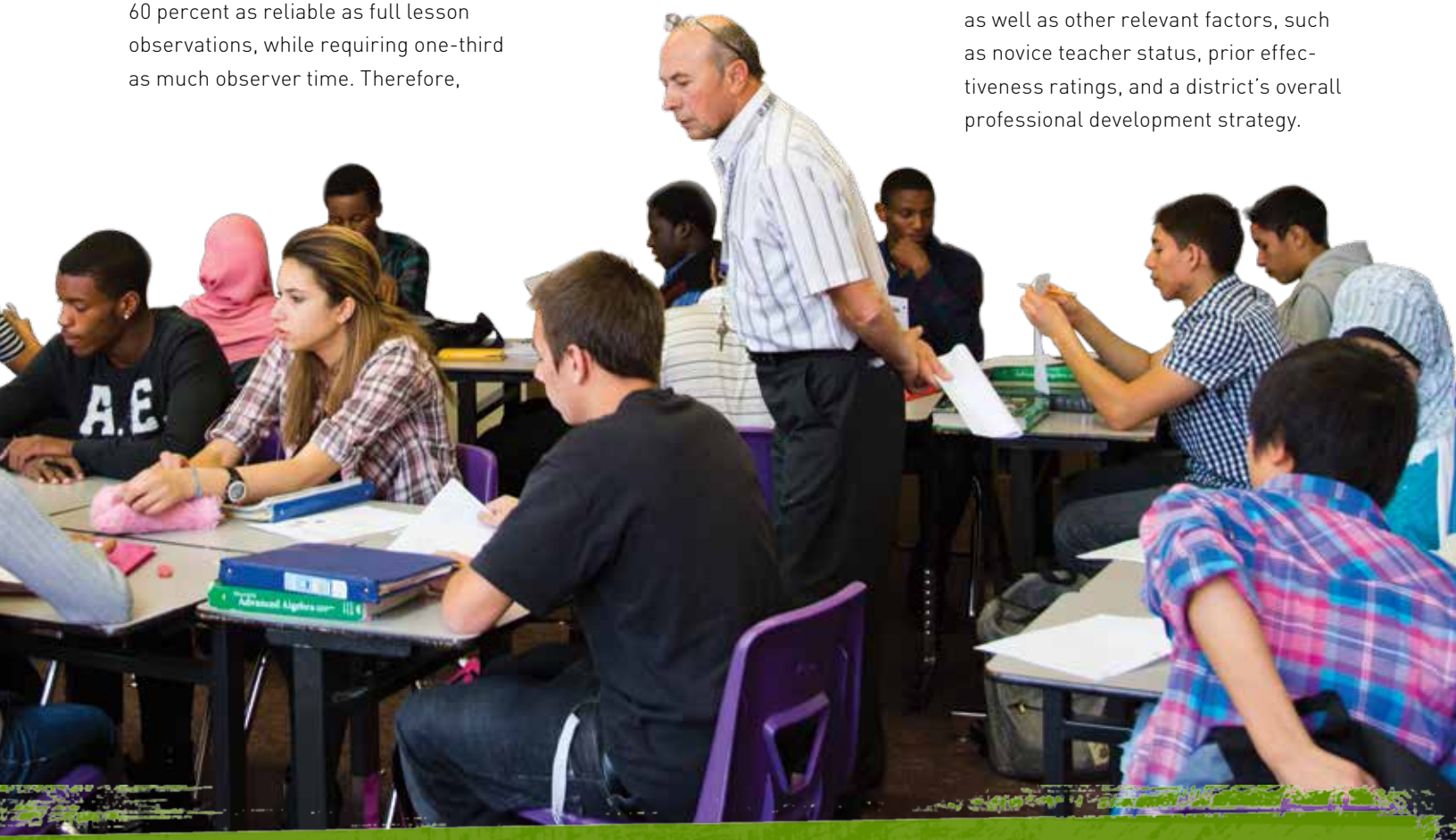
Our analysis from Hillsborough County showed observations based on the first 15 minutes of lessons were about 60 percent as reliable as full lesson observations, while requiring one-third as much observer time. Therefore, one way to increase reliability is to expose a given teacher's practice to multiple perspectives. Having three different observers each observe for 15 minutes may be a more economical way to improve reliability than having one additional observer sit in for 45 minutes. Our results also suggest that it is important to have at least one or two full-length observations, given that some aspects of teaching scored on the Framework for Teaching (Danielson's instrument) were frequently not observed during the first 15 minutes of class.

"Although administrators gave higher scores to their own teachers, their rankings of their own teachers were similar to those produced by external observers and administrators from other schools."

Together, these results provide a range of scenarios for achieving reliable classroom observations. There is a point where both additional observers and additional observations do little to reduce error. Reliability above 0.65 can be achieved with several configurations (see **Figure 5**).

Implications for Districts

Ultimately, districts must decide how to allocate time and resources to classroom observations. The answers to the questions of how many lessons, of what duration, and conducted by whom are informed by reliability considerations, as well as other relevant factors, such as novice teacher status, prior effectiveness ratings, and a district's overall professional development strategy.



What We Know Now

In three years we have learned a lot about how multiple measures can identify effective teaching and the contribution that teachers can make to student learning. The goal is for such measures to inform state and district efforts to support improvements in teaching to benefit all students. Many of these lessons have already been put into practice as school systems eagerly seek out evidence-based guidance. Only a few years ago the norm for teacher evaluation was to assign "satisfactory" ratings to nearly all teachers evaluated while providing virtually no useful information to improve practice.[10] Among the significant lessons learned through the MET project and the work of its partners:

- **Student perception surveys and classroom observations can provide meaningful feedback to teachers.** They also can help system leaders prioritize their investments in professional development to target the biggest gaps between teachers' actual practice and the expectations for effective teaching.
- **Implementing specific procedures in evaluation systems can increase trust in the data and the results.** These include rigorous training and certification of observers; observation of multiple lessons by different observers; and in the case of student surveys, the assurance of student confidentiality.
- **Each measure adds something of value.** Classroom observations provide rich feedback on practice. Student perception surveys provide a reliable indicator of the learning environment and give voice to the intended beneficiaries of instruction. Student learning gains (adjusted to account for differences among students) can help identify groups of teachers who, by virtue of their instruction, are helping students learn more.
- **A balanced approach is most sensible when assigning weights to form a composite measure.** Compared with schemes that heavily weight one measure, those that assign 33 percent to 50 percent of the weight to student achievement gains achieve more consistency, avoid the risk of encouraging too narrow a focus on any one aspect of teaching, and can support a broader range of learning objectives than measured by a single test.
- **There is great potential in using video for teacher feedback and for the training and assessment of observers.** The advances made in this technology have been significant, resulting in lower costs, greater ease of use, and better quality.

The Work Ahead

As we move forward, MET project teachers are supporting the transition from research to practice. More than 300 teachers are helping the project build a video library of practice for use in professional development. They will record more than 50 lessons each by the end of this school year and make these lessons available to states, school districts, and other organizations committed to improving effective teaching.

This will allow countless educators to analyze instruction and see examples of great teaching in action.

Furthermore, the unprecedented data collected by the MET project over the past three years are being made available to the larger research community to carry out additional analyses, which will increase knowledge of what constitutes effective teaching and how to support it. MET project partners already are tapping those data for new studies on observer training, combining student surveys and observations, and other practical concerns. Finally, commercially available video-based tools for observer training and certification now exist using the lessons learned from the MET project's studies.

Many of the future lessons regarding teacher feedback and evaluation systems must necessarily come from the field, as states and districts innovate, assess the results, and make needed adjustments. This will be a significant undertaking, as systems work to better support great teaching. Thanks to the hard work of MET project partners, we have a solid foundation on which to build.

"Many of the future lessons regarding teacher feedback and evaluation systems must necessarily come from the field, as states and districts innovate, assess the results, and make needed adjustments. This will be a significant undertaking, as systems work to better support great teaching."

Endnotes

1. The lead authors of this brief are Steven Cantrell, Chief Research Officer at the Bill & Melinda Gates Foundation, and Thomas J. Kane, Professor of Education and Economics at the Harvard Graduate School of Education and principal investigator of the Measures of Effective Teaching (MET) project. Lead authors of the related research papers are Thomas J. Kane (Harvard), Daniel F. McCaffrey (RAND), and Douglas O. Staiger (Dartmouth). Essential support came from Jeff Archer, Sarah Buhayar, Alejandro Ganimian, Andrew Ho, Kerri Kerr, Erin McGoldrick, and David Parker. KSA-Plus Communications provided design and editorial assistance.

2. This section summarizes the analyses and key findings from the research report *Have We Identified Effective Teachers?* by Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. Readers who want to review the full set of findings can download that report at www.metproject.org.

3. As expected, not every student on a randomly assigned roster stayed in the classroom of the intended teacher. Fortunately, we could track those students. We estimated the effects of teachers on student achievement using a statistical technique commonly used in randomized trials called "instrumental variables."

4. These predictions, as well as the average achievement outcomes, are reported relative to the average among participating teachers in the same school, grade, and subject.

5. Readers may notice that some of the differences in Figure 2 are smaller than the differences reported in earlier MET reports. Due to non-compliance—students not remaining with their randomly assigned teacher—only about 30 percent of the randomly assigned difference in teacher effectiveness translated into differences in the effectiveness of students' actual teacher. The estimates in Figure 2 are adjusted for non-compliance. If all the students had remained with their randomly assigned teachers, we would have predicted impacts roughly three times as big. Our results imply that, without non-compliance, we would have expected to see differences just as large as included in earlier reports.

6. Other researchers have studied natural movements of teachers between schools (as opposed to randomly assigned transfers) and found no evidence of bias in estimated teacher effectiveness between schools. See Raj Chetty, John Friedman, and Jonah E. Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," working paper no. 17699, National Bureau of Economic Research, December 2011.

7. The findings highlighted in this summary and the technical details of the methods that produced them are explained in detail in the research paper "A Composite Estimator of Effective Teaching," by Kata Mihaly, Daniel McCaffrey, Douglas O. Staiger, and J.R. Lockwood. A copy may be found at www.metproject.org.

8. Different student assessments, observation protocols, and student survey instruments would likely yield somewhat different amounts of reliability and accuracy. Moreover, measures used for evaluation may produce different results than seen in the MET project, which attached no stakes to the measures it administered in the classrooms of its volunteer teachers.

9. This section summarizes key analyses and findings from the report *The Reliability of Classroom Observations by School Personnel* by Andrew D. Ho and Thomas J. Kane. Readers who want to review the full set of findings and methods for the analyses can download that report at www.metproject.org. The MET project acknowledges the hard work of Danni Greenberg Resnick and David Steele, of the Hillsborough County Public Schools, and the work of the teachers, administrators, and peer observers who participated in this study.

10. Weisburg, D. et al. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* Brooklyn: New Teacher Project.

Bill & Melinda Gates Foundation

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—especially those with the fewest resources—have access to the opportunities they need to succeed in school and life. Based in Seattle, Washington, the foundation is led by CEO Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

For more information on the U.S. Program, which works primarily to improve high school and postsecondary education, please visit www.gatesfoundation.org.

©2013 Bill & Melinda Gates Foundation. All Rights Reserved.
Bill & Melinda Gates Foundation is a registered trademark in the United States and other countries.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org