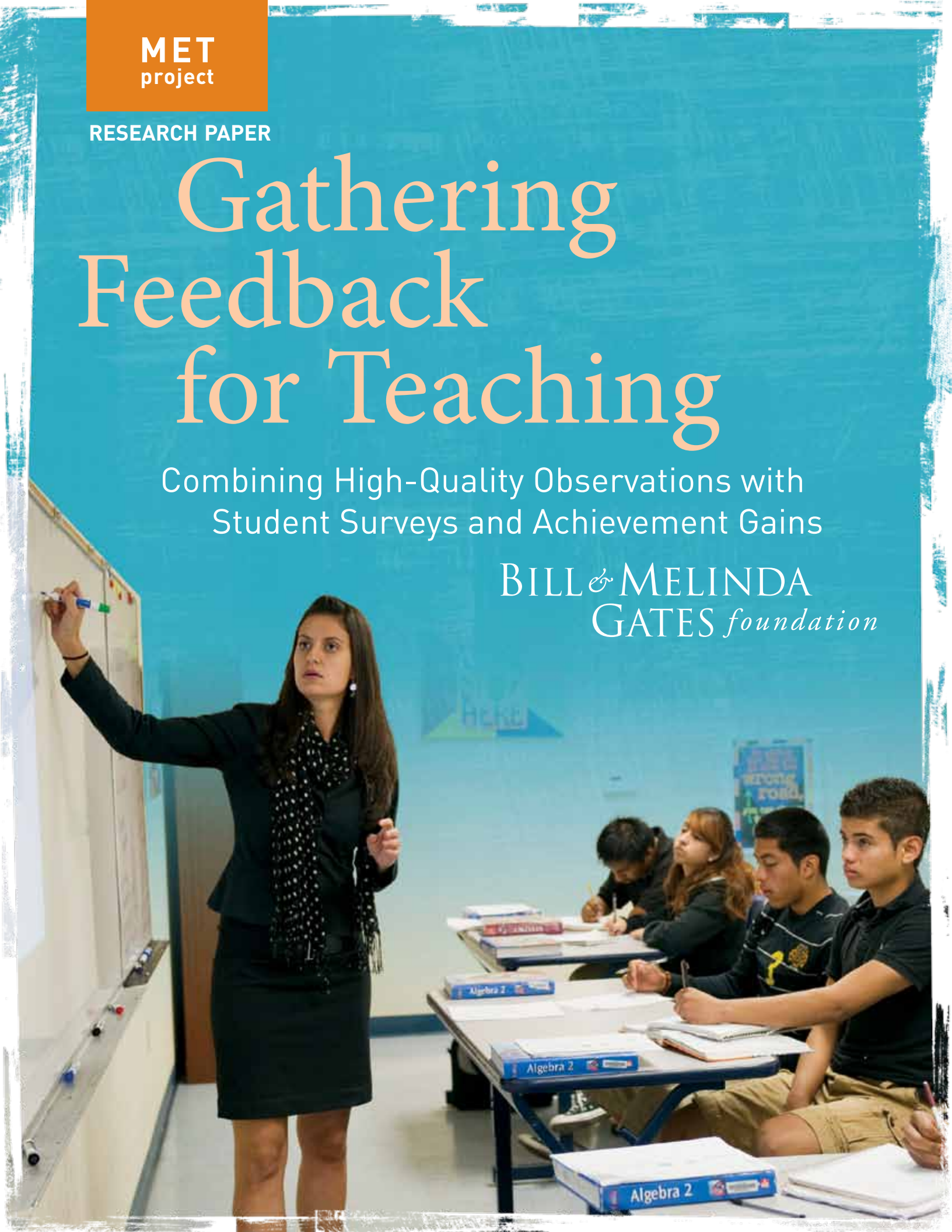
MET project

RESEARCH PAPER

Gathering Feedback for Teaching

Combining High-Quality Observations with Student Surveys and Achievement Gains

BILL & MELINDA GATES foundation



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.

January 2012

©2012 Bill & Melinda Gates Foundation. All Rights Reserved.
Bill & Melinda Gates Foundation is a registered trademark in the United States and other countries.

About This Report

This report presents an in-depth discussion of the analytical methods and findings from the Measures of Effective Teaching (MET) project's analysis of classroom observations.[1] A nontechnical companion report describes implications for policymakers and practitioners.

Together, these two documents on classroom observations represent the second pair of publications from the MET project. In December 2010, the project released its initial analysis of measures of student perceptions and student achievement in *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.* Two more reports are planned for mid-2012: one on the implications of assigning weights to different measures; another testing the validity of teacher effectiveness measures following random assignment of students to teachers.

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. Lead research and organizational partners include:

- Mark Atkinson, Teachscape
- Joan Auchter, National Board for Professional Teaching Standards
- Nancy Caldwell, Westat
- Charlotte Danielson, The Danielson Group
- Ron Ferguson, Harvard University
- Drew Gitomer, Rutgers University
- Dan Goldhaber, University of Washington
- Pam Grossman, Stanford University
- Heather Hill, Harvard University
- Eric Hirsch, New Teacher Center
- Sabrina Laine, American Institutes for Research
- Michael Marder, University of Texas
- Dan McCaffrey, RAND
- Catherine McClellan, Educational Testing Service
- Denis Newman, Empirical Education
- Roy Pea, Stanford University
- Raymond Pecheone, Stanford University
- Geoffrey Phelps, Educational Testing Service
- Robert Pianta, University of Virginia
- Morgan Polikoff, University of Southern California

1 Lead authors of this report were Thomas J. Kane, Deputy Director of Research and Data at the Bill & Melinda Gates Foundation and Professor of Education and Economics at the Harvard Graduate School of Education, and Douglas O. Staiger, Professor of Economics at Dartmouth College. Key analyses were conducted by Dan McCaffrey of the Rand Corporation. Steve Cantrell, Jeff Archer, Sarah Buhayar, Kerri Kerr, Todd Kawakita, and David Parker assisted on project direction, data collection, writing, and analysis.

- Rob Ramsdell, Cambridge Education
- Steve Raudenbush, University of Chicago
- Brian Rowan, University of Michigan
- Doug Staiger, Dartmouth College
- John Winn, National Math and Science Initiative

Acknowledgments

In addition to MET project partners who reviewed early drafts of this report, we would like to thank the following external experts who read and provided written feedback: Anthony Bryk, Andrew Ho, Bob Linn, Susan Moore Johnson, and Jonah Rockoff. The lead authors accept full responsibility for any remaining errors in the analysis.

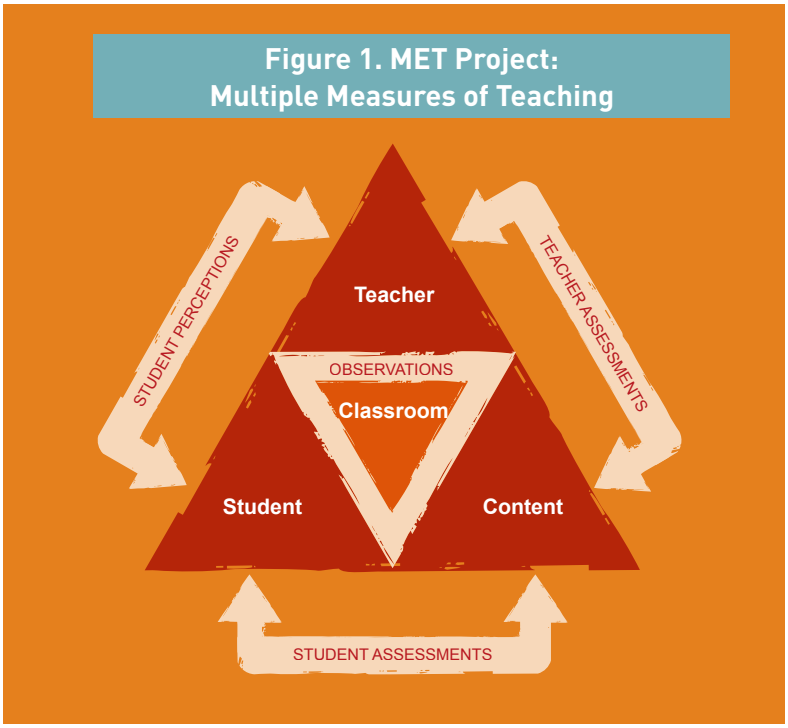
We want to express particular gratitude to the nearly 3,000 teachers who as MET project volunteers opened up their practice to help the project gain insights that can strengthen the teaching profession and improve outcomes for students.

Introduction and Executive Summary

There is a growing consensus that teacher evaluation in the United States is fundamentally broken. Few would argue that a system that tells 98 percent of teachers they are "satisfactory" benefits anyone—including teachers. The nation's collective failure to invest in high-quality professional feedback to teachers is inconsistent with decades of research reporting large disparities in student learning gains in different teachers' classrooms (even within the same schools). The quality of instruction matters. And our schools pay too little attention to it.

Many states and school districts are looking to reinvent the way they do teacher evaluation and feedback, and they want better tools. With the help of nearly 3,000 teacher-volunteers, the Measures of Effective Teaching (MET) project is evaluating alternative ways to provide valid and reliable feedback to teachers for professional development and improvement (see **Figure 1**). In our first report in December 2010, we reported on the potential value of feedback from student surveys. In this report, we focus on the value of classroom observations.

Figure 1. MET Project: Multiple Measures of Teaching



The MET project is unique in several ways:

- **Scale:** As far as we know, this is the largest study of instructional practice and its relationship to student outcomes.
- **Range of indicators:** We compare many different instruments for classroom observation and evaluate different combinations of measures, including student feedback and student achievement gains.
- **Range of student outcomes:** Rather than focus solely on student scores on state tests, our data provide a unique window into other outcomes, including an open-ended literacy assessment and an assessment of conceptual understanding in mathematics, as well as student-reported levels of effort and enjoyment being in a teacher's class.
- **Random assignment:** During the second year of the study, teachers agreed to be randomly assigned to classes. In a mid-2012 report, we will test whether these measures of effective teaching are able to predict teachers' student outcomes following random assignment.

In this report, we test five different approaches to classroom observations. Each observation instrument is designed to do two things: (1) focus an observer's attention on specific aspects of teaching practice and (2) establish common evidentiary standards for each level of practice. Ideally, an observation instrument should create a common vocabulary for pursuing a shared vision of effective instruction.

We investigate the following five instruments in this report:[2]

- **Framework for Teaching** (or **FFT,** developed by Charlotte Danielson of the Danielson Group),
- **Classroom Assessment Scoring System** (or **CLASS,** developed by Robert Pianta, Karen La Paro, and Bridget Hamre at the University of Virginia),
- **Protocol for Language Arts Teaching Observations** (or **PLATO,** developed by Pam Grossman at Stanford University),
- **Mathematical Quality of Instruction** (or **MQI,** developed by Heather Hill of Harvard University), and
- **UTeach Teacher Observation Protocol** (or **UTOP,** developed by Michael Marder and Candace Walkington at the University of Texas-Austin).

These instruments are not checklists, focusing on easy-to-measure but trivial aspects of practice. They require training and judgment on the part of observers. For example, FFT emphasizes the importance of a teacher's questioning techniques. The instrument describes low-quality questioning as occurring when a teacher asks a series of questions requiring one-word, yes/no answers, poses such questions in rapid succession with little wait time, and involves only a few students in the classroom. In contrast, the instrument defines high-quality questioning as requiring students to reveal their reasoning, giving students time to consider their answers, and involving many students in the class.

Most educators would agree that questioning technique is an important competency in a teacher's repertoire. But they would also want to know, "Is it possible to describe high- and low-quality questioning techniques sufficiently clearly so that observers can be trained to recognize strong questioning skills?"; "Would different observers come to similar judgments?"; and, ultimately, "Are those judgments related to student outcomes measured in different ways?" Those are the questions we set out to answer in this report.

TWO CRITERIA

Specifically, we compare the classroom observation instruments using two criteria:

First, for each of the observation instruments, we estimated the *reliability* with which trained observers were able to characterize persistent aspects of each teacher's practice, using thousands of hours of lessons collected for this project. Reliability is simply the proportion of the variance in instrument scores reflecting consistent differences in practice between individual teachers (as opposed to variation attributable to the particular observer, or group of students being taught, or even the particular lesson delivered). Reliability is important because without it classroom observations will paint an inaccurate portrait of teachers' practice. We estimated the reliability for each instrument based on thousands of hours of digital video, with the same teachers working with different sections of students, delivering different lessons, with scores provided by different raters.

2 One of our partners in this research, the National Board for Professional Teaching Standards, has provided data for those applying for certification from the MET project districts. We also are investigating a sixth observation instrument, Quality Science Teaching (or QST, developed by Raymond Pecheone and Susan Schultz at Stanford, for assessing high school science instruction). Results from both of these will be included in our final report in mid-2012.

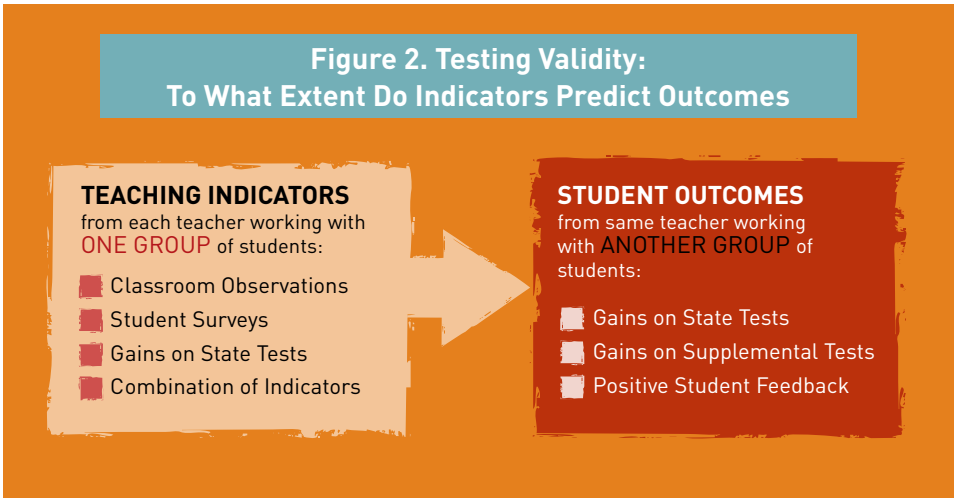
Much of the prior research with the observation instruments has been done by the instrument developers themselves, with scoring performed by small research teams. In such circumstances, it can be hard to distinguish between the power of the instrument and the special expertise of the instrument developers themselves to discern effective teaching. However, to be useful to schools, the instruments need to be transferable. We don't just want to know whether a small group of experts can distinguish between effective and ineffective instruction; we want to know whether a larger group of observers with little special expertise beyond a background in teaching can be trained to look for the same competencies. Therefore, the Educational Testing Service (ETS, one of the project partners) mounted a large recruitment effort and trained more than 900 observers to score classroom videos using the various instruments.

Second, we report the association between the observations and a range of different student outcomes: achievement gains on state tests and on other, more cognitively challenging assessments, as well as on student-reported effort and enjoyment while in class. Each of the observation instruments represents a set of hypotheses about what "effective teaching" looks like. By applying all five instruments to a given set of videos, we essentially test these hypotheses by investigating whether the teachers with higher scores on each instrument also had higher student outcomes. Of course, observation scores will not be perfectly aligned with student achievement gains and other outcomes. There may be some teaching competencies that affect students in ways we are not measuring. Moreover, since when comparing any two measures there will be measurement error in both, some teachers with high scores on classroom observations will have low student achievement gains—and vice versa. Nonetheless, if the observation instruments are identifying instructional practices that actually help students learn, we ought to see that achievement gains and the scores on an instrument are aligned on average.

We tried to isolate the effects of teaching from any pre-existing student characteristics that could affect student outcomes at the end of the year. For example, when calculating each student's achievement gain on the state and supplemental tests, as well as on student-reported outcomes, we controlled statistically for the individual student's characteristics (including prior state test scores) and the mean characteristics of all the students in each classroom (to account for peer effects). This approach is sometimes referred to as "value-added" modeling.

Obviously, other factors, besides prior achievement and student demographic characteristics, may influence student achievement. Therefore, we collected data from *more than one* group of students for each teacher (either from another course section or school year). Because they were watching videos, raters could notice other student traits and behaviors (such as unusual attentiveness or misbehavior) that we are not controlling for in the value-added measure. If so, the observed student achievement gain and the classroom observation score may both be influenced by these unmeasured student background characteristics. As a precaution, we used the videos of a teacher's practice with one group of students to compare against the achievement gains for *another* group of students, in a different section or academic year taught by the same teacher (see **Figure 2**).[3]

3 Of course, if multiple groups of students taught by the same teacher share the same unmeasured trait, the above approach would still be biased. Ultimately, the only way to test for this is by randomly assigning students to teachers, which we did for the second year of the study. We will be reporting the results from that random assignment validation in mid-2012.



FINDINGS

The analysis in this report is based on the practice of 1,333 teachers from the following districts: Charlotte-Mecklenburg, N.C.; Dallas; Denver; Hillsborough Co., Fla.; New York City; and Memphis. This is the subset of MET project volunteers who taught math or English language arts (ELA) in grades 4 through 8 and who agreed to participate in random assignment during year 2 of the project.[4] For this report, MET project raters scored 7,491 videos of lessons at least three times: once each with the cross-subject instruments, CLASS and FFT, and a third time on the subject-specific instruments, either MQI or PLATO. We scored a subset of 1,000 math videos a fourth time with the UTOP instrument.[5] In addition, we incorporated data on state test scores, supplemental tests, and student surveys from more than 44,500 students.

Five findings stood out:

1. All five observation instruments were positively associated with student achievement gains.

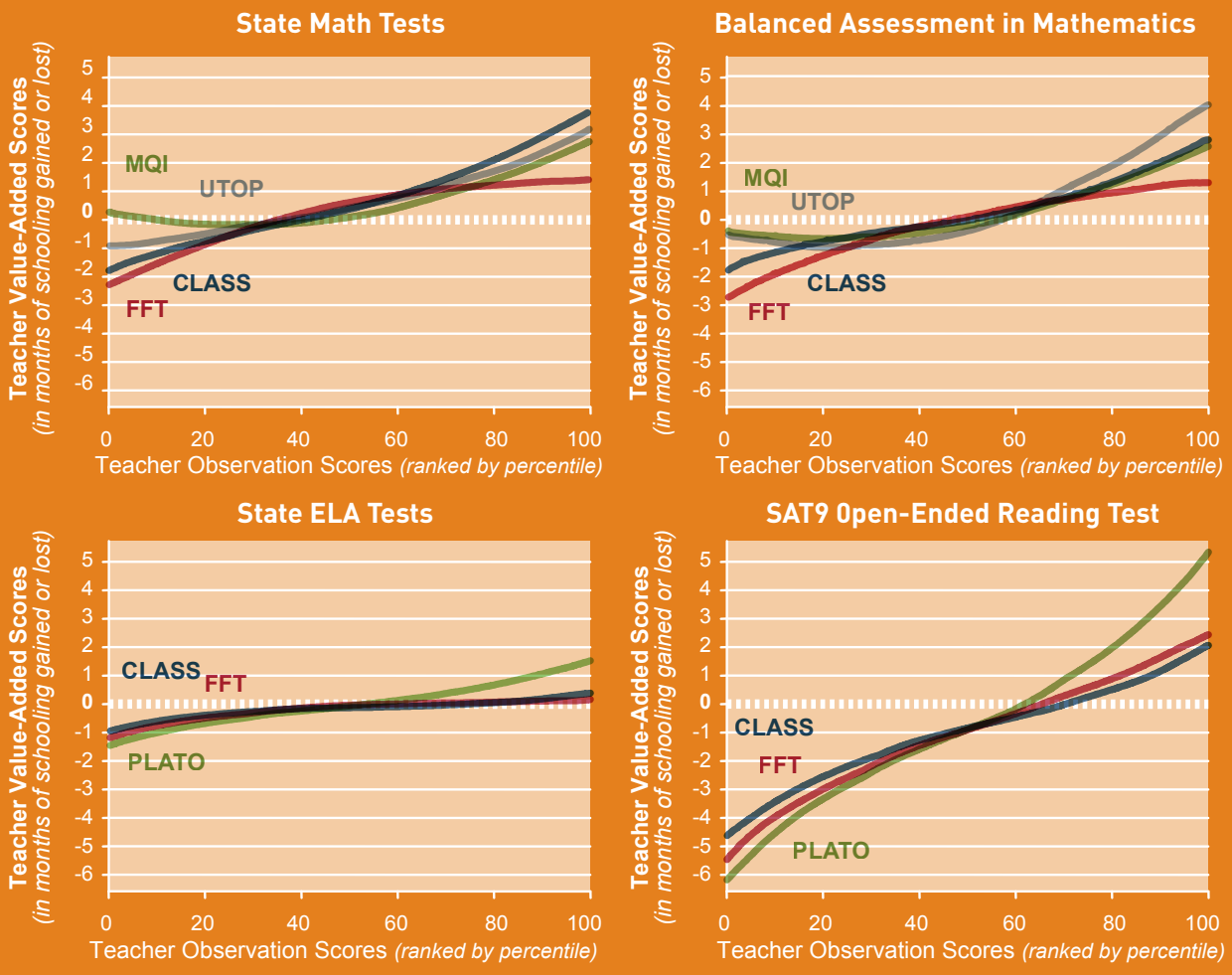
Ultimately, the goal is to use classroom observations to help teachers improve student outcomes. A classroom observation system that bears no relationship to student outcomes will be of little help in doing so.

To investigate validity, we use two types of achievement gains. Given concerns over the quality of current state tests, many have worried that those teachers who achieve large gains on the state tests are simply coaching children to take tests, not teaching the underlying concepts. For this reason, the MET project administered two assessments to supplement the state test results: the Balanced Assessment in Mathematics (BAM) and the open-ended version of the Stanford 9 (SAT9 OE) reading test. The BAM test was designed to measure students' conceptual understanding of math topics. The open-ended version of the Stanford 9 provides a series of reading passages (like the existing state ELA tests) but, unlike most state tests, asks students to write short-answer responses to questions testing their comprehension, rather than asking them to choose an answer in multiple-choice format. (For examples of these two assessments, see the appendix of our prior report: *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.*)

4 With a few exceptions described in Appendix Table 1, we did not score the videos from the MET project teachers who did not participate in random assignment.

5 As explained further in the following pages, scoring of UTOP was managed by the National Math and Science Initiative (NMSI), whereas ETS managed scoring of the other instruments.

Figure 3. Teachers with Higher Observation Scores Had Students Who Learned More



NOTES: Value-added estimated in student-level standard deviation units and converted to months of schooling using conversion factor of 0.25 standard deviations = 9 months of schooling. Slopes were calculated as running regressions. Teachers' value-added scores and observation scores were from working with different groups of students.

Figure 3 provides a graphical summary of the relationships between teachers' classroom observation scores and student achievement gains on four different types of assessments: state tests in math and ELA, the BAM test, and the SAT9 OE reading assessment. On the horizontal axis is the percentile rank of teacher scores on each of the classroom observation instruments.[6] On the vertical axis is the average value-added of a second group of students taught by the same teacher in 2009–10.[7] Value-added was estimated in units of student-level standard deviations within grade and subject for each test.[8] However, for the figures, we translated the

6 For each instrument, the measure is the average score on two different lessons, each scored by a different rater from one course section in 2009–10. If scoring more lessons produced higher reliability, the relationships would be expected to be steeper.

7 As reported in the paper, we also look at the relationship between observation scores in 2009–10 and their students' achievement gains in 2008–09, with similar results.

8 Some readers may prefer the standard-deviation metric. To convert from months to standard deviations, they need only divide the months of schooling by 36. Because we did not have access to separate conversion factors for each type of assessment and in order to make it easier for readers to convert between standard deviations and months, we used the same conversion factor for all assessments.

value-added measures into "months of schooling" using a conversion factor of 0.25 standard deviations per 9 months of schooling.[9] The top and bottom panels of Figure 3 are reported for math and ELA respectively. (We use the same scale on both, so that the slopes are comparable.[10]) Three facts are evident:

- All of the instruments were related to student achievement gains in math and ELA. For example, students in classes taught by teachers in the bottom quartile (below the 25th percentile) in their classroom observation scores using FFT, CLASS, or UTOP *fell behind* comparable students with comparable peers by roughly 1 month of schooling in math. In contrast, students with teachers with observation scores in the top quartile (above the 75th percentile) *moved ahead* of comparable students by 1.5 months (and even more for those at the top end of the UTOP scale).

- The differences were roughly half as large on state ELA tests as in math.[11] Other researchers have reported similar findings: Teachers have smaller effects on student scores on state assessments of ELA than on state math assessments. However, this may reflect the nature of the current state ELA assessments, not whether teachers of literacy have a smaller influence than teachers of math.

- As reported in the bottom right panel, there is a stronger association between classroom observation scores and student achievement gains on the open-ended reading assessment than on the state ELA tests. In fact, the association is comparable to that observed in mathematics. Therefore, it is on the state ELA tests where the relationships between value-added and observations are weaker, not on all literacy tests. This may be because beyond the early grades, teachers are no longer teaching reading comprehension alone but are also teaching writing skills.

2. Reliably characterizing a teacher's practice requires averaging scores over multiple observations.

Even with systematic training and certification of observers, the MET project needed to combine scores from multiple raters and multiple lessons to achieve high levels of reliability. A teacher's score varied considerably from lesson to lesson, as well as from observer to observer. For four out of five observation instruments, we could achieve reliabilities in the neighborhood of 0.65 only by scoring four different lessons, each by a different observer. Many have commented on the volatility (i.e., lack of reliability) in value-added measures. However, in our study, a single observation by a single observer was often more volatile (and, therefore, less reliable) than value-added: in our study, single observations produced reliabilities ranging from 0.14 to 0.37. By comparison, researchers typically report reliability of value-added measures between 0.30 to 0.50.

9 Using the vertical scale scores from the Stanford 9 norm sample as well as age cut-offs for school enrollment in Los Angeles, Kane (2004) infers that 9 months of schooling is associated with a 0.25 standard deviation gain in performance. Neal and Johnson (1996) use variation in educational attainment associated with quarter of birth to generate a similar estimate of the impact of later schooling on scores on the Armed Forces Qualification Test.

10 We do not force any of these relationships to follow a straight line; we allow the slope to be steeper in the middle or at the high and low ends. We produced these graphs using a running regression-line smoother to allow for a very general type of nonlinear relationship.

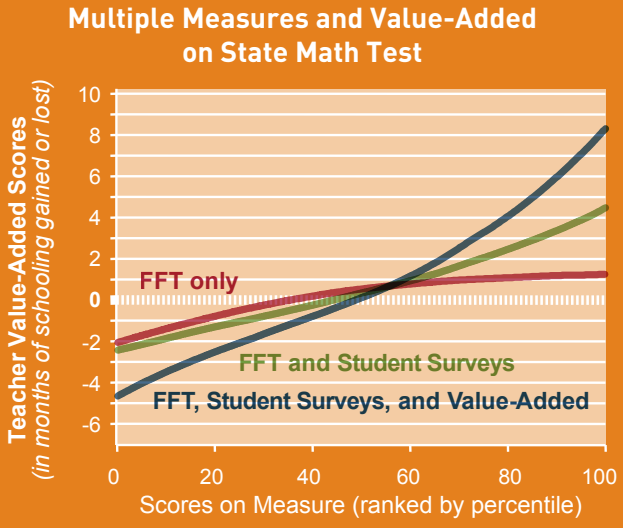
11 A one percentile point difference in observation scores had larger implications for student achievement at the top and bottom of the distribution of scores than in the middle. But these relationships were roughly linear in the scales of the instruments themselves. A one percentile point difference represents a larger absolute difference in the scaled scores at the bottom and top than at the middle of the distribution.

We caution against extrapolating from our results too literally and concluding that a state or district would require four observations, each by a different observer, to obtain similar reliabilities. The context of our study was unique in several obvious ways: (1) our observers were trained and required to demonstrate their ability to score accurately before they could begin scoring, (2) our observers had no personal relationship with the teachers being observed, (3) our observers were watching digital video and were not present in the classrooms, (4) there were no stakes attached to the scores for teachers in our study (or for the observers). Nevertheless, whether they can achieve reliability with less than four observations or not, school systems are likely to face the same two challenges: instructional practice for a given teacher varies from lesson to lesson and, even with training, the instruments all require rater judgment, which is rarely unanimous.

3. Combining observation scores with evidence of student achievement gains and student feedback improved predictive power and reliability.

Figure 4 compares the predictive power of three different approaches to measuring effective teaching: (a) using classroom observation scores alone, (b) combining classroom observation scores with student feedback, and (c) combining observation scores with student feedback and the achievement gains a teacher saw with a different group of students. (The student feedback was collected using the Tripod student perception survey, developed by Ron Ferguson at Harvard University, which was the focus of our last report). We refer to this third measure, which combines classroom observations with student feedback and student achievement gains, as the "combined measure." Figure 4 uses FFT as the observational instrument and gains on state math tests, but results look similar for all of the other instruments. Two facts are evident:

Figure 4. Combining Measures Added Predictive Power



NOTES: Value-added estimated in student-level standard deviation units and converted to months of schooling using conversion factor of 0.25 standard deviations = 9 months of schooling. Slopes were calculated as running regressions. Teachers' value-added scores and scores of measures were from working with different groups of students. Combined measure was created with equal weights.

- When moving from the "observation only" measure to the "observation + student feedback," the difference in achievement gain between the top and bottom quartile teachers increases in math, from 2.6 months to 4.8 months. Although not shown in Figure 4, the gain from adding the student survey data was similar for ELA.

- Predictive power increased further when a teacher's value-added measure was added to the set of predictors. (The three measures were standardized to have equal variance and equally weighted when combined.) For example, when teachers were ranked on the combined measure, the difference between having a top- and bottom-quartile teacher was nearly 8 months in math and 2.5 months in ELA.

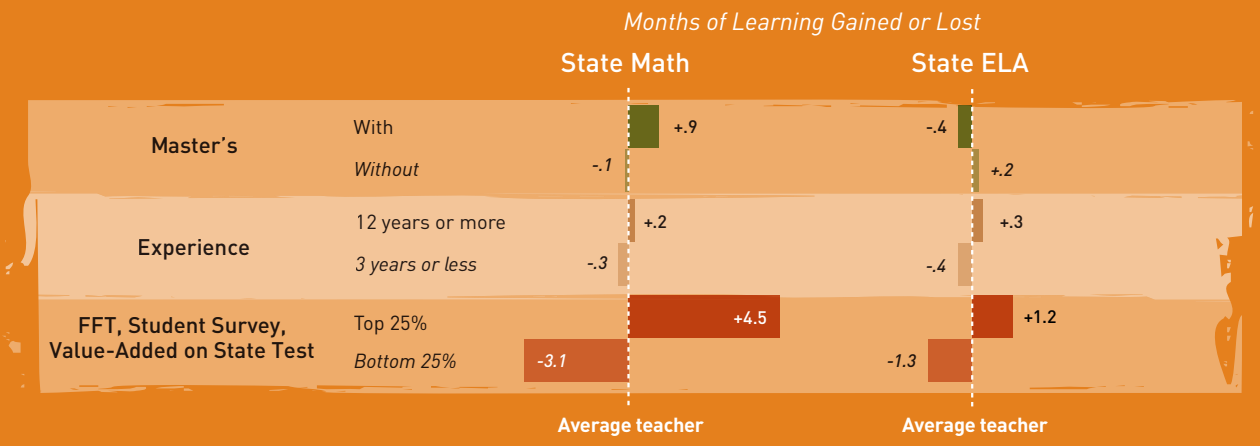
Put simply, adding evidence of student achievement gains improves the chance of identifying teachers who see larger student achievement gains with other students. These are very large differences, roughly equal to one-fifth of the black-white achievement gap in the country and nearly equal to the gains students make during the typical school year.

Combining measures has another advantage: higher reliability. For example, the reliability of a single classroom value-added measure was 0.48 in math and 0.20 in ELA; the reliability of two classroom observations done by different observers in one class ranged from 0.24 to 0.53 for the different classroom observation measures; the reliability of student feedback across different classrooms was highest, 0.66. However, depending on the classroom observation instrument used, the reliability of the combined measure ranged from 0.55 to 0.67 in math and 0.51 to 0.54 in ELA.[12] In other words, combining multiple measures led not only to higher predictive power but greater reliability (more stability) as well.

4. In contrast to teaching experience and graduate degrees, the combined measure identifies teachers with larger gains on the state tests.

No measure is perfect. But better information (even if imperfect) should allow for better decisions. We compared the predictive power of the combined measure to two alternatives: teaching experience and graduate degrees. Currently, these are the primary determinants of teacher pay and promotion in school districts around the country. For example, when school districts faced fiscal challenges recently, seniority was often the sole criterion used to determine layoffs. To achieve comparable levels of selectivity, we compared those in

Figure 5. Combined Measures Better Identified Effective Teaching on State Tests than Master's or Experience



NOTES: Value-added estimated in student-level standard deviation units and converted to months of schooling using conversion factor of 0.25 standard deviations = 9 months of schooling. Teachers' value-added scores and scores of measures were from working with different groups of students. Combined measure was created with equal weights. Differences between the top and bottom 25 percent on the combined measure are significant at the 0.001 level. None of the differences for master's and experience is signifcant at the 0.05 level.

12 These are single course section reliabilities. All would be higher if averaged over multiple course sections in a single year or multiple years.

the top 25 percent of teaching experience (roughly 12+ years of teaching) to those in the bottom 25 percent of teaching experience (0 to 3 years of teaching). About 35 percent of the MET project teachers had master's degrees or higher.

The bar chart in **Figure 5** reports the average gain in math and ELA—relative to comparable students with comparable peers—for all three comparisons. To be consistent with other figures (which use outcomes that we collected only in 2009–10), we construct the combined measure in Figure 5 using data from one course section, and we evaluate its ability to predict student achievement gains in another course section in 2009–10.[13] Compared to teachers with fewer than three years of experience, teachers with 12 or more years of experience had students with slightly higher achievement gains in math (0.5 months higher) and slightly higher achievement gains in ELA (0.7 months higher). The difference for those with master's degrees was also small: 1 month higher in math and actually 0.6 months lower in ELA compared to those teachers without master's. These are much smaller than the differences noted using the combined measure: nearly 8 months on the state math tests and 2.5 months on the state ELA tests.

5. Teachers with strong performance on the combined measure also performed well on other student outcomes:

- Their students showed larger performance gains on tests of conceptual understanding in mathematics and a literacy test requiring short-answer responses.
- Their students reported higher levels of effort and greater enjoyment in class.

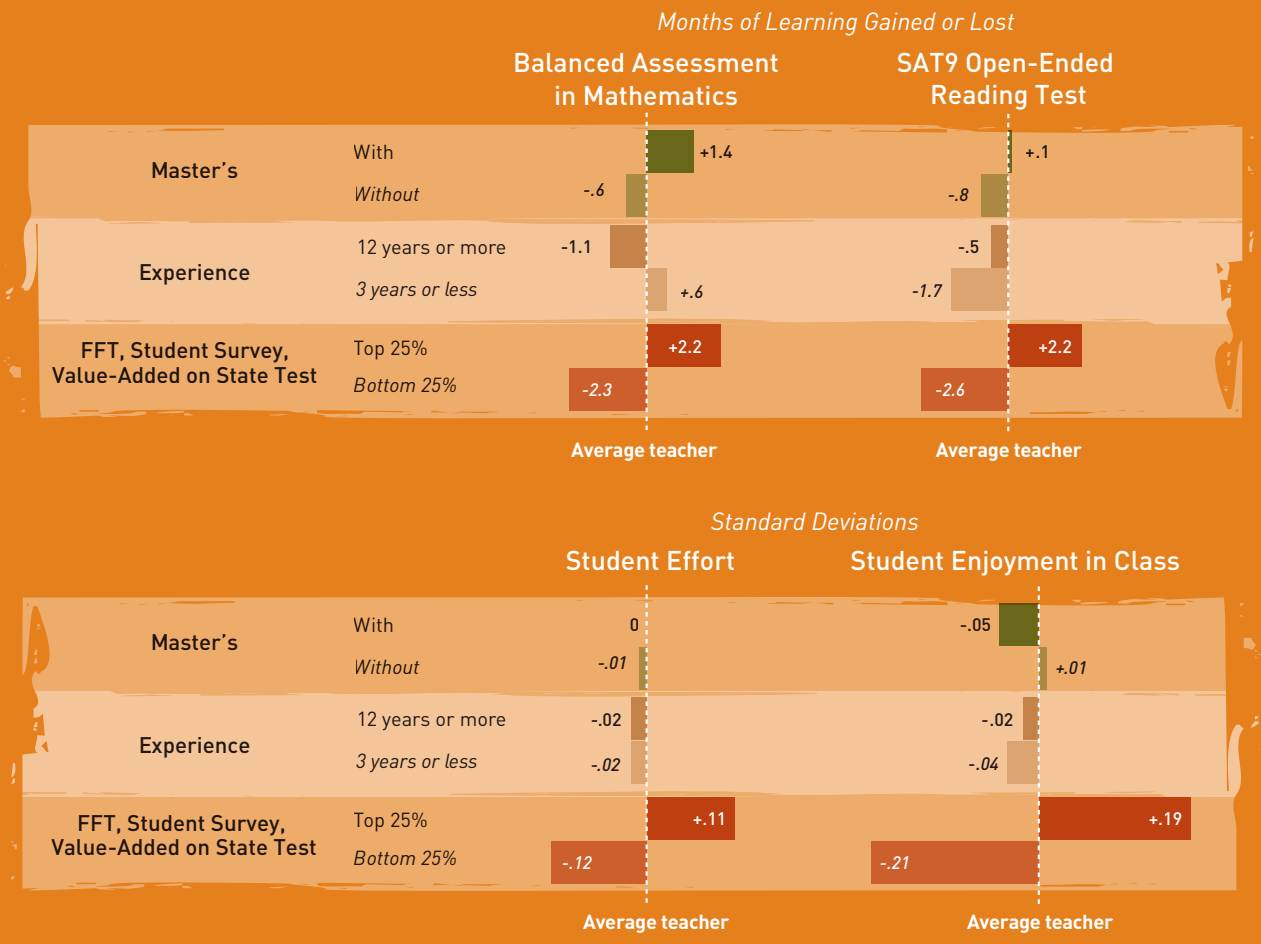
The format of **Figure 6** is similar to Figure 5. We constructed the combined measure using student achievement gains on state tests (as well as the Tripod student survey and FFT scores). But rather than report the average student performance gain on state tests, Figure 6 reports the difference in student achievement gains on the supplemental assessments between teachers in the top and bottom quartiles. (The results from the supplemental assessments were not used in the combined measure used to create the chart.) Students saw gains on the supplemental assessments that were 4.5 months of schooling greater on the test of math concepts (BAM) and 4.8 months greater on the SAT9 OE reading test, compared with teachers in the bottom quartile.

In contrast, those teachers with master's degrees saw average student gains of about 2 months on BAM and about 0.9 months on SAT9 OE relative to those without master's degrees. The most experienced teachers saw average student gains of about 1.7 months fewer in BAM and about 1.2 months more in SAT9 OE compared to the least experienced teachers.

The state test gains are not perfectly correlated with the supplemental test gains. Obviously, the best way to measure improvement on the type of skills measured in the supplemental assessments would be to change the state assessments to include such skills. However, it would be incorrect to assume that the teachers whose students had large gains on the current math tests are not teaching children conceptual understanding in mathematics or the ability to write short-answer responses in ELA. Those with larger gains on the state tests

13 However, the results are similar if we were to use all data from 2009–10 to predict results from 2008–09.

Figure 6. Combined Measures Better Identified Effective Teaching on Supplemental Tests and Better Predicted Positive Student Feedback than Master's or Experience



NOTES: Value-added estimated in student-level standard deviation units and converted to months of schooling using conversion factor of 0.25 standard deviations = 9 months of schooling. Teachers' value-added scores and scores of measures were from working with different groups of students. Combined measure was created with equal weights. Combined measure quartiles in the bottom chart are made up of teachers at the top and bottom on either state math or state ELA tests. Differences between the top and bottom 25 percent on the combined measure are significant at the 0.001 level. None of the differences for master's and experience is significant at the 0.05 level, except differences between master's and without master's for BAM (p = 0.03).

did have higher gains on the supplemental tests on average.[14] Moreover, there's no evidence to suggest that making decisions on the basis of a teacher's experience and possession of a graduate degree would be better aligned with the types of skills measured by the supplemental tests.

But parents care about other outcomes, such as whether their children are developing a positive emotional attachment to school and are engaged in their learning. Parents would not want teachers to emphasize practices that improve achievement but lead children to dread school or to work less hard. To investigate the extent to which our measures of teaching predict indications of student effort and students' enjoyment of being in a

14 The underlying teacher effect on state tests and supplemental tests were moderately correlated in our data—0.45 correlation between state math and BAM, and 0.46 correlation between state ELA and SAT9 OE. As we emphasize later, this moderate correlation implies that teachers whose students ranked highly on high-stakes state assessments will also tend to perform well on alternative cognitively demanding (and low-stakes) assessments. But this moderate correlation also implies that value-added estimates based on state tests are in part driven by aspects of teacher performance that are specific to the state test and that may not generalize to other student outcomes of interest. Thus, as we discuss later, there will be benefits to evaluating teachers based on multiple measures to minimize the influence of teacher performance that is idiosyncratic to any specific measure.

teacher's class, we identified related items from the Tripod student survey, such as "When doing schoolwork for this class, I try to learn as much as I can and I don't worry how long it takes" and "This class is a happy place for me to be."

Figure 6 reports differences in student outcomes on these two metrics. The teachers who ranked highly on the combined measure, including FFT, had students in other classes who reported significantly more positive emotional attachment to being in a teacher's class (0.4 student-level standard deviations more than teachers ranked in the bottom quartile), and they reported higher levels of effort (0.23 standard deviations more than the bottom quartile). In contrast, teachers' experience and whether or not they had a graduate degree did not help students on these nonacademic student outcomes. (In fact, those with master's degrees had students with lower reported levels of attachment than those without them, though the differences were small.)

IMPLICATIONS

The evidence in the report has at least three implications for districts and states implementing new teacher evaluation systems:

1. **Achieving high levels of reliability of classroom observations will require several quality assurances: observer training and certification; system-level "audits" using a second set of impartial observers; and use of multiple observations whenever stakes are high.**

Not all decisions require high levels of reliability. Measures could be used many different ways: promotion decisions, retention decisions, compensation decisions, or low-stakes feedback intended to support improvement. Different uses necessitate different evidentiary standards and different levels of reliability (there is no uniform standard that applies to any envisioned use).[15] For instance, when making a high-stakes decision regarding promotion or retention, a district should seek high levels of reliability. In contrast, when simply making suggestions for experienced teachers, the requisite level of reliability would be lower. The former would require multiple observations, whereas a single observation might be enough for the latter. As we noted earlier, our findings cannot tell how many observations would be required in a real setting because of the unique context of our study.

Even so, we feel confident offering the following guidance to policymakers and practitioners for ensuring high-quality classroom observations:

- Observers should be trained and expected to **demonstrate their ability to use an observation instrument accurately *before*** beginning high-stakes classroom observations.

Good training is necessary but likely insufficient to ensure accuracy. At the end of training, anyone who will be asked to assess instruction and provide feedback based on classroom observations should demonstrate a minimum level of accuracy in rating lessons before they actually evaluate teachers using an instrument. This could be done through the rating of pre-scored videos or through simultaneous live observations that allow trainees to compare notes with expert observers.

- When high stakes are involved, teachers should be observed **during *more than one* lesson.**

15 For more on a framework for thinking about false positives and false negatives, please see Glazerman et al. (2010).

Our findings suggest that no one lesson provides a complete picture of a teacher's practice. Different lessons may call for different competencies. Any teacher could have an off day. This problem cannot be solved with high inter-rater reliability alone, since there seems to be variation in a teacher's practice from lesson to lesson. To achieve adequate reliability, we found it necessary to average scores across multiple lessons.

- Districts and states should systematically **track the reliability of their classroom observation procedures.**

One way to monitor reliability is to have a subset of teachers observed by impartial observers (who come from outside the teachers' school and have no personal relationship to the teachers) and compare the impartial scores with the official scores.[16]

2. Evaluation systems should include multiple measures, not just observations or value-added alone.

Combining measures offers three advantages: greater predictive power (slightly better than student achievement gains alone, but significantly better than observations alone), greater reliability (especially when student feedback or multiple observation scores are included), and the potential for diagnostic insight to allow teachers to improve their practice (which cannot be provided by student achievement gains alone).

When the goal is to identify teachers likely to have large gains on the state test with other students, the best single predictor is a similar measure of achievement gains (value-added) in another course section or academic year. While there is a small improvement in predictive power when adding observations and student feedback, it is only a modest gain.

But there are other benefits to having multiple measures. For instance, because the reliability of student feedback was higher than the other measures in our study, incorporating student feedback generally raised reliability (lowered volatility) when it was included in a combined measure. Although an individual student may have a less sophisticated understanding of effective instruction than a trained observer, student feedback has two other advantages that contribute to reliability: students see the teacher all year (and, therefore, are less susceptible to lesson to lesson variation), and the measures are averaged over 20 to 75 students, rather than 1 or 2 observers. When multiple classroom observations from more than one lesson could be averaged together, these also produced higher reliability (but a single observation is unlikely to help much and could actually lower reliability).

3. The true promise of classroom observations is the potential to identify strengths and address specific weaknesses in teachers' practice.

Admittedly, there is little evidence that untargeted professional development (that is not based on an individualized assessment of a teacher's strengths and weaknesses) has a large impact. But a recent study by Allen et al. (2011) suggests that individualized coaching for secondary school teachers using the CLASS instrument

16 An indirect method that could also be helpful would be to monitor the relationship between classroom observation scores and other measures, such as student achievement gains. If observation scores are based on factors other than effective instruction—such as inadequate training of observers, idiosyncratic judgments by supervisors, or personal relationships between supervisors and teachers (positive or negative)—one would expect them to lose their alignment with other outcomes, such as student achievement gains. One danger of this approach is that observers would begin to base their observation scores partially on value-added to avoid raising flags.

leads to substantial improvements in student achievement. Another recent paper by Taylor and Tyler (2011) found that providing clear feedback to teachers using the FFT instrument led to substantial improvements in student achievement gains in teachers' classrooms (even without a targeted professional development effort).

When value-added data are available, classroom observations add little to the ability to predict value-added gains with other groups of students. Moreover, classroom observations are less reliable than student feedback, unless many different observations are added together. Therefore, the real potential of classroom observations is their usefulness for diagnosis and development of instructional practice. School systems should be looking for ways to use classroom observations for developmental purposes. In the near future, we will need more powerful examples of how to use individualized feedback to facilitate development as well as evidence of impact.

WHAT'S IN THIS REPORT

The sections are organized as follows:

- **Data Collection and Descriptive Statistics** (pp. 16–27). In this section, we describe the sample for our study and the instruments we investigated. We explain our process for capturing instruction in the form of digital video recordings and for having those videos scored by observers trained in each of the five observation instruments. We also report on the resulting distribution of scores we found for each instrument.

- **The Principal Components of Instruction** (pp. 28–33). In this section, we use our data to ask: To what extent do the discrete competencies defined in these instruments appear together, in clusters of competencies, or independently? The answer is critical to being able to identify the importance of any individual competency. When competencies usually appear together in clusters, it is impossible to say which individual competencies matter most. We also investigate the extent to which the different cross-subject and single-subject instruments seem to be measuring similar or different things.

- **The Reliability of Classroom Observations** (pp. 34–40). In this section, we describe the different sources of variance in observation results, including variance by observer, by lesson, and of course, by teacher. We then estimate the incremental gain in reliability for measuring consistent aspects of a teacher's practice from using different numbers of observations and observers.

- **Validating Classroom Observations with Student Achievement Gains** (pp. 41–59). In this section, we explain our method for testing the alignment of the observation results with student outcomes. We investigate the extent to which those with higher scores on each of the instruments also have larger student achievement gains. We compare the predictive validity of observation instruments when taken alone, as well as in combination with student feedback and with student achievement gains with a different group of students.

Throughout the report, we remind readers of the special conditions under which we conducted our research and caution against expectations that a state or district would find the same results. Nonetheless, the sections listed above offer general guiding principles and recommendations based on the patterns we see in the data.

In the last pages we look ahead to our next set of analyses and explain the focus of the MET project's final report in mid-2012. Finally, we urge states, districts, and others to apply these same tests to their own data on their own measures of teacher performance and to use the findings to inform continuous improvement of these measures.

Data Collection and Descriptive Statistics

Teachers in the study arranged to have themselves recorded four to eight times over the course of the year. These videos were then scored by trained raters using each of the observation instruments. In addition, we administered confidential student perception surveys in each teacher's class using the Tripod survey instrument that was the subject of our last report. We also collected state test score results for the students of the teachers in the study. Finally, to those students we administered supplemental assessments that included open-ended items and that are designed to test higher-order thinking.

The analysis for this report is based on data from the classrooms of 1,333 teachers in grades 4 through 8. (As reported in Appendix Table 1, this subset excludes the MET project teachers working in grade 9 and the teachers who were not randomized in the second year of the study.) This includes videos of more than 7,491 lessons, roughly half of which captured math lessons and half captured ELA. All lessons were scored on the two cross-subject instruments (FFT and CLASS) and on either MQI or PLATO, depending on the content. A subset of 1,000 math lessons were also scored on UTOP. In total, this produced more than 22,000 observation scores for our analysis. In addition, data from perception surveys, state tests, and supplemental assessments were collected on more than 44,500 students.

Table 1 shows the mean characteristics of students in the MET project classrooms.

Table 1. Student Characteristics

	PERCENTAGE
Hispanic	31
Black/American Indian	33
White/Asian	34
Gifted	11
Male	50
SPED	8
ELL	13
Subsidized Lunch	56

Note: This table reports the mean characteristics of students who had state assessment data in math or ELA. New York City did not report gifted student status and Charlotte did not report subsidized lunch status. These percentages for those characteristics are thus based on the districts that did report.

Table 2 shows the mean teacher characteristics of MET project volunteer teachers compared to non-MET project teachers in the same districts. The MET project teachers were quite similar to their peers in terms of demographics, educational attainment, and average experience. (Later, in Table 12, we compare the MET project teachers to their peers in terms of value-added in math and ELA.)

Table 2. Characteristics of MET Project Volunteers vs. Other Teachers in Districts

		RACE/ETHNICITY				YEARS OF TEACHING			
		% WHITE	% BLACK	% HISPANIC	% OTHER	TOTAL	IN DISTRICT	% FEMALE	% MASTER'S DEGREE OR HIGHER
All Districts	MET	56.8	35.4	5.6	2.2	10.3	7.2	83.3	36.4
	Non-MET	59.0	26.9	10.8	3.3	11.2	8.5	82.8	33.3
Charlotte Mecklenberg	MET	71.2	25.5	3.3	0.0			84.5	35.3
	Non-MET	66.7	30.6	2.7	0.0			85.6	40.0
Dallas	MET	32.0	54.5	7.9	5.6	9.7	7.4	74.6	30.0
	Non-MET	33.3	54.9	8.9	2.9	11.5	9.0	78.7	32.4
Denver	MET	92.2	3.9	3.9	0.0	7.2	5.2	72.6	58.0
	Non-MET	75.8	7.4	15.8	1.0	8.8	7.1	79.2	46.2
Hillsborough	MET	76.4	13.0	8.8	1.9	10.5	8.9	85.2	16.0
	Non-MET	72.3	16.1	9.7	1.9	11.4	9.2	85.2	18.3
Memphis	MET	21.8	77.6	0.3	0.3	10.9	4.5	88.7	30.0
	Non-MET	23.5	76.0	0.1	0.4	11.8	5.1	85.5	32.4
New York City	MET	63.7	25.0	7.0	4.4		7.4	83.0	
	Non-MET	58.3	24.3	13.0	4.4		8.5	82.0	

Note: Based on only 2010 data and teachers of grades 4–8 in math or ELA. Shaded cells indicate that the variable is not available for a particular district.

THE TRIPOD STUDENT PERCEPTION SURVEY

The student perception survey used in the MET project is based on more than a decade of work by the Tripod Project for School Improvement. Tripod was created by a team led by Ronald F. Ferguson of Harvard University.[17] The student survey asks students their level of agreement with a series of statements related to different aspects of classroom climate and instruction (e.g., "If you don't understand something, my teacher explains it another way.") These statements are organized under seven constructs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate.

To ensure honest feedback, the Tripod surveys were administered confidentially, so that teachers could not tie responses back to students.

SUPPLEMENTAL ASSESSMENTS OF STUDENT LEARNING

To collect achievement data in addition to that provided by state tests, the MET project administered two supplemental assessments to grade 4–8 students in participating classes in spring 2010: the Balanced Assessment in Mathematics (BAM) and SAT9 OE reading assessment. These were chosen because they both include cognitively demanding content and are reasonably aligned to the state curricula followed by the six districts. BAM includes four to five tasks, requires 50–60 minutes to complete, and measures higher-order reasoning skills using question formats that are quite different from those in most state mathematics achievement tests.

17 For more information on Tripod and MET project findings on the survey, see *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.*

The SAT9 OE reading assessment contains nine open-ended tasks and takes 50 minutes to complete. The primary difference between the SAT9 OE and traditional state reading tests is its exclusive use of open-ended items tied to extended reading passages. Students are required not only to answer the question but also to explain their answers.

VIDEO CAPTURE OF CLASSROOM INSTRUCTION

To record lessons, the MET project used a panoramic camera that simultaneously captures two views from a fixed position: a 360-degree perspective and a higher-resolution stationary view of the classroom whiteboard. Microphones on the camera and worn by the teacher picked up the teacher's voice and whole-group discussion. Because the rig was stationary, no camera crew was needed. Participating teachers worked with site-based project coordinators to schedule recordings. Guidelines stipulated that at least two recorded lessons for each teacher involve a topic from a list of subject-specific "focal topics." Examples of focal topics included, for 5th grade math, adding and subtracting fractions and for 7–9th grade ELA, writing about literature.

THE MET PROJECT'S OBSERVATION INSTRUMENTS

The observation instruments used to assess video-recorded lessons in the MET project share a number of features. Developed based on literature reviews, expert opinion, and empirical evidence, they divide teaching and learning into discrete aspects of practice, and they break each into three to seven performance levels. The instruments are designed for making expert inferences about the quality of particular aspects of the teaching observed, as opposed to simply checking for whether certain practices are present.

Most of the instruments have been subjected to tests of validity and reliability by their developers, though the extent of the research varies greatly depending largely on how long the instruments have existed. While there is overlap in the aspects of instruction they value, each instrument is rooted in a particular vision of quality instruction. These visions are reflected in the various elements of teaching that the instruments define. (Instrument developers alternately call these elements "components," "dimensions," "sections," and "indicators"; for the sake of consistency in discussing multiple instruments, we use the term "competency" in this report.) **Table 3** presents an overview of the origin and structure of the instruments used in the MET project. **Table 4** presents all of the competencies for all of the instruments as used in the MET project.

Table 3. Highlights of Observation Instruments as Used in the MET Project[18]

Instrument	Lead Developer	Origin	Instructional Approach	Grades	Subjects	Structure, as used in MET project	Scoring
Framework for Teaching (FFT)	Charlotte Danielson, the Danielson Group	Developed as an outgrowth of Danielson's work on Educational Testing Service's PRAXIS III assessment for state licensing of new teachers	Grounded in a "constructivist" view of student learning, with emphasis on intellectual engagement	K–12	All academic subjects	2 domains, subdivided into 8 components *Note: Includes 2 additional domains—"planning and preparation" and "professional responsibilities"—that could not be observed in the videos*	4-point scale
Classroom Assessment Scoring System (CLASS)	Robert Pianta, University of Virginia	Initially developed as a tool for research on early childhood development	Focus on interactions between students and teachers as the primary mechanism of student learning	K–12 *Note: 2 versions used in MET project: Upper Elementary and Secondary*	All academic subjects	3 domains of teacher-student interactions subdivided into 11 "dimensions," plus a fourth domain on student engagement	7-point scale; scores assigned based on alignment with anchor descriptions at "high," "mid," and "low"
Protocol for Language Arts Teaching Observations (PLATO) *Note: Version used for MET project, "Plato Prime"*	Pam Grossman, Stanford University	Created as part of a research study on ELA-focused classroom practices at middle grades that differentiate more and less effective teachers	Emphasis on instructional scaffolding through teacher modeling, explicit teaching of ELA strategies, and guided practice	4–9	English language arts	6 elements of ELA instruction *Note: Full instrument includes 13 elements*	4-point scale
Mathematical Quality of Instruction (MQI) *Note: Version used for MET project, "MQI Lite"*	Heather Hill with colleagues at Harvard and University of Michigan	Designed as tool for capturing classroom practices associated with written tests of math teaching knowledge	Instrument stresses teacher accuracy with content and meaning-focused instruction	K–9	Math	6 elements of math instruction *Note: Full version includes scores for 24 subelements*	3-point scale
UTeach Teacher Observation Protocol (UTOP)	UTeach teacher preparation program at University of Texas-Austin	Observation tool created by model program for preparing math and science majors to become teachers	Designed to value different modes of instruction, from inquiry-based to direct	K–college	Math, science, and computers *Note: Used in MET project for math*	4 sections, subdivided into 22 total subsections	5-point scale

18 The MET project also is investigating a sixth instrument—Quality Science Teaching (QST)—developed at Stanford to assess biology instruction. However, data on QST were not yet available for this report. Analysis of the instrument will be included in future MET project publications.

Table 4. Domains and Scored Competencies from Observation Instruments as Used in MET Project Analysis[19]

CLASS *Cross subject*	*Emotional Support*				*Classroom Organization*		
	Positive climate	Negative climate	Teacher sensitivity	Regard for student perspectives	Behavior management	Productivity	Instructional learning formats
	Instructional Support				*Student Engagement*		
	Quality of feedback	Content understanding	Analysis and problem solving	Instructional dialogue	Student Engagement		
Framework for Teaching *Cross subject*	*Classroom Environment*						
	Creating an environment of respect and rapport		Establishing a culture of learning		Managing classroom procedures		Managing student behavior
	Instruction						
	Communicating with students		Using questioning and discussion techniques		Engaging students in learning		Using assessments in instruction
PLATO Prime *ELA specific*	Behavior Management			Time Management			Explicit Strategy use and instruction
	Modeling			Classroom Discourse			Intellectual Challenge
MQI Lite *Math specific*	Classroom Work Connected to Math			Richness of Mathematics			Working with Students and Mathematics
	Errors and Imprecision			Student Participation in Meaning Making and Reasoning			Explicitness and Thoroughness in Content Presentation
UTOP *Math and science specific*	*Classroom Environment*						
	Student generation of ideas/questions	Collegiality among students		Intellectual engagement with key ideas	Majority of students on-task	Classroom management	Attention to access, equity, and diversity
	Lesson Structure						
	Lesson organization	Structures for student engagement		Investigation/problem-based approach	Appropriate resources		Critical and reflective on practice
	Mathematics Content						
	Significance of content	Explicitness of importance	Teacher knowledge and fluency	Accuracy of teacher written content	Use of abstraction and representation	Connections to other disciplines/ math areas	Relevance to history, current events
	Implementation						
	Questioning strategies		Use of formative assessments		Involvement of all students		Allocation of time

Note: For MQI, the table doesn't reflect scores given for two overall assessments of instructional quality: "Overall Mathematical Quality of Instruction" and "Lesson-Based Guess at Knowledge of Teaching."

19 Competencies shown may differ from those in typical version of the instrument due to adjustments made for MET project scoring.

RECRUITING RATERS

Two MET project partners, ETS and Teachscape, jointly managed the recruitment and training of raters and lesson scoring. (The one exception was the UTOP instrument, which was managed by the National Math and Science Initiative [NMSI]).

Instrument developers set minimum expectations for the education level and teaching experience of raters. All raters held a bachelor's degree, and a majority (about 70 percent across most instruments) held higher degrees. While some raters were currently enrolled in teacher preparation programs, the vast majority (more than 75 percent) had six or more years of teaching experience.

Table 5. Rater Education

	HIGHEST DEGREE ATTAINED		
INSTRUMENT	% BACHELOR'S	% MASTER'S	% PH.D.
FFT	28	63	7
CLASS	30	62	6
PLATO	23	71	6
MQI	31	55	14
UTOP	43	54	3

Note: Totals may not be 100 percent due to missing data.

Table 6. Rater Teaching Experience

	TEACHING STATUS (PERCENTAGE)			YEARS OF EXPERIENCE (PERCENTAGE, AMONG CURRENT/FORMER TEACHERS)			
	CURRENT TEACHER	FORMER TEACHER	NEVER TAUGHT	1–2 YEARS	3–5 YEARS	6–10 YEARS	10+ YEARS
FFT	44	47	9	4	10	23	63
CLASS	41	45	14	9	18	25	48
PLATO	40	58	2	0	17	26	57
MQI	29	46	25	12	20	12	56
UTOP	N/A	N/A	0	0	5	19	76
All	**40**	**48**	**12**	**7**	**16**	**25**	**52**

N/A = Not available.

ETS recruited observers using a range of online methods: postings on the ETS web site; postings on educational professional web sites (e.g., National Council of Teachers of Mathematics, National Council of Teachers of English); emails to ETS scorers, such as those scoring Advanced Placement exams; and postings on Facebook. The vast majority of applicants came from the postings on the ETS web site, followed by the professional web sites.

RATER TRAINING

Depending on the instrument, rater training required between 17 and 25 hours to complete. Training for the four instruments (other than UTOP) was conducted via online, self-directed modules. Raters for UTOP were trained using a combination of in-person and online sessions. Training for all of the instruments included:

- Discussion of the instrument, its competencies, and its performance levels;
- Video examples of teaching for each competency at each performance level;
- Practice scoring videos, with feedback from trainers; and
- Techniques for minimizing rater bias.

RATER CERTIFICATION

At the end of their training, raters were required to rate a number of pre-scored videos and achieve a minimum level of agreement with the expert scores. Raters who failed certification after one attempt were directed to review the training material. Those who failed after a second attempt were deemed ineligible to score for the MET project (see **Figure 7**).The pass rate for raters averaged 77 percent across instruments and ranged from 56 percent (MQI) to 83 percent (FFT).

Table 7. Rater Certification Criteria and Pass Rates[20]

Instrument	Certification Criteria	Pass Rate after 2 tries
FFT	• At least 50 percent exact match of correct scores • No more than 25 percent ratings "discrepant" (scores two or more off from the correct score on the instrument's four-point scale)	83%
CLASS	• 70 percent of scores had to either exactly match the correct score or be "adjacent" to the correct score on the instrument's seven-point scale	82%
MQI	• MQI defines its certification standard not in terms of percentage matches, but in terms of the maximum allowable difference between averages of raters' scores and correct scores	56%
PLATO	• At least 70 percent exact match with correct scores on the instrument's four-point scale • No more than 10 percent rater scores discrepant from correct scores	71%
	Pass Rate across All Instruments	**77%**

ONGOING RATER MONITORING AND SUPPORT

The MET project also monitored rater accuracy on an ongoing basis. At the start of each shift, raters had to pass a calibration exercise, scoring a smaller set of pre-scored videos. Raters who failed to pass calibration after two tries (about 9 percent per day) were not permitted to score videos that day. Throughout their work with the project, raters received additional training and guidance from their "scoring leader"—an expert scorer responsible for managing and supervising a group of raters.

20 UTOP training, managed by NMSI, did not include such a certification process. Instead, UTOP raters trained for the MET project scored three videos and normed their understandings in group discussions at the end of in-person training sessions. Because it was managed by NMSI and not ETS, scoring for UTOP differed in four important ways from the other four instruments: UTOP raters received in-person training; UTOP raters viewed entire lessons, whereas those using the other four instruments viewed the first 30 minutes of each lesson; the UTOP developers recruited and trained their own raters, whereas ETS recruited and trained the raters for the other instruments; and approximately one-third of lessons rated on UTOP were double scored, compared with 5 percent for the others. Arguably, these differences may have boosted the reliability of UTOP scores relative to the other four instruments.

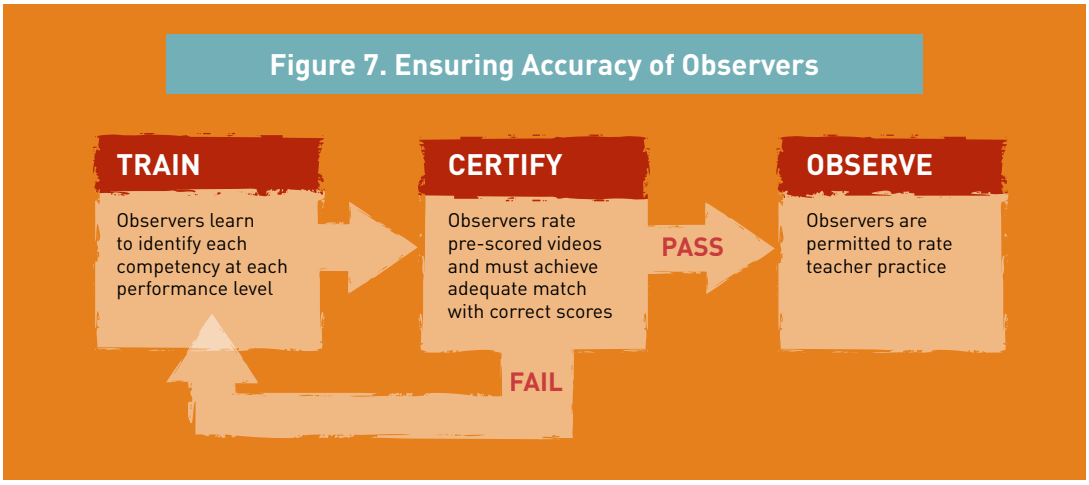


Table 8. Calibration Pass/Fail Rates by Instrument

Instrument	% times passed on first try	% times passed only after two tries	% times failed after two tries
FFT	84	10	6
CLASS-Secondary	61	22	17
CLASS-Upper Elementary	67	18	15
MQI	94	5	1
PLATO	91	7	1
Across All Instruments	**77**	**14**	**9**

In addition, pre-scored videos were interspersed within the unscored videos assigned to each rater (although raters were not told which videos were pre-scored and which were unscored). Scoring leaders were provided reports on the rates of agreement their raters were able to achieve with those videos. Scoring leaders were asked to work with raters who frequently submitted discrepant scores on the pre-scored videos. Double scoring, in which the same lesson was scored by two raters, also served as a form of quality control.

DISTRIBUTION OF OBSERVATION SCORES

The MET project data provide a unique glimpse into the classroom practices of urban teachers. Raters scored lessons in segments that varied in length depending on the instrument (15 minutes for CLASS, for instance, and 7.5 minutes for MQI), and these data reflect that unit of analysis. For ease of exposition, "lesson segment" and "lesson" are used interchangeably in the discussion below. **Figure 8** (pp. 26–27) shows the distribution of scores, by each competency, on each instrument for lessons taught by MET project teachers.

CLASS

Where practice was strongest: Raters scored the lessons taught by MET project teachers highest on two related dimensions, *behavior management* and *productivity.* The distribution of practice was nearly identical across these competencies, with more than two-thirds of scores among the top two (6 or 7) performance levels and 85 percent of scores among the top three performance levels (5 or higher). The vast majority of teachers also avoided creating a negative climate for their students, with fewer than 1 percent of scores among the top three (most negative) performance levels.

Where practice was weakest: Lessons taught by MET project teachers were scored lowest on the most complex aspects of teaching: *analysis and problem solving, regard for student perspectives, quality of feedback, instructional dialogue,* and *content understanding.* Fewer than 10 percent of scores were in the top two performance levels. Of these competencies, *analysis and problem solving* was least frequently observed, with just 20 percent of lesson segments scored in the top four (of seven) performance levels.

Framework for Teaching

Where practice was strongest: Lessons taught by MET project teachers scored highest on *managing student behavior, creating an environment of respect and rapport,* and *engaging students in learning.* The distribution of practice was nearly identical across these competencies, with more than two-thirds of lesson segments scored as "proficient" or "distinguished." The first two competencies address the necessity of establishing an environment where learning can take place, and *engaging students in learning* addresses the materials, activities, and structure of the lesson.

Where practice was weakest: Lessons taught by MET project teachers scored lowest on: *communicating with students, using questioning and discussion techniques,* and *using assessments in instruction.* Fewer than half of all lesson segments were rated "proficient" or better. Of these competencies, *communicating with students* was least frequently observed, with just 30 percent scoring "proficient" or better. This is puzzling and requires some unpacking of what is required to score proficient on *communicating with students.* This label is perhaps misleadingly simple, yet the practices it entails are quite complicated, including clear presentation of content, culturally and developmentally appropriate communication (both required to score "proficient"), and identification of student misconceptions (required to score "distinguished").

PLATO Prime

Where practice was strongest: Lessons taught by MET project teachers scored highest on *behavior management* and *time management.*

Where practice was weakest: For *intellectual challenge* and *classroom discourse* slightly more than one-third of the lesson segments were scored at proficient or above. Even fewer were rated proficient or better for *explicit strategy use and instruction* and *modeling.* For these two areas, more than half of the lessons rated were unsatisfactory.

MQI Lite

Where practice was strongest: Nearly all of the mathematics lessons (93.4 percent) contain *classroom work connected to mathematics* and very few of these lesson segments were wrought with mathematical *errors and imprecision.*

Where practice was weakest: Mathematics lessons were scored very low for three of the competencies: *working with students and mathematics, richness of mathematics,* and *student participation in meaning making and reasoning.* For these competencies, about 1 percent of lessons achieved the highest rating and more than 70 percent of lessons received the lowest rating.

UTOP

Where practice was strongest: Mathematics lessons scored on UTOP rated highly on competencies related to managing the classroom environment, with 70 percent or more receiving scores of three or above on the instrument's five-point scale for *majority of students on-task* and *classroom management.* High scores also went for competencies related to teachers' presentation of material, with about 85 percent of lessons scoring a three or above on *accuracy of teacher written content,* 80 percent scoring similarly for *use of abstraction and representation,* and more than 70 percent for *teacher knowledge and fluency.*

Where practice was weakest: Many of the lowest scores given for UTOP competences related to skills associated with engaging students in rigorous instruction. Only about 35 percent of lessons scored a three or above on *intellectual engagement with key ideas.* The same was true for about 30 percent of lessons on questioning strategies and only about 15 percent on *investigation/problem-based approach.*

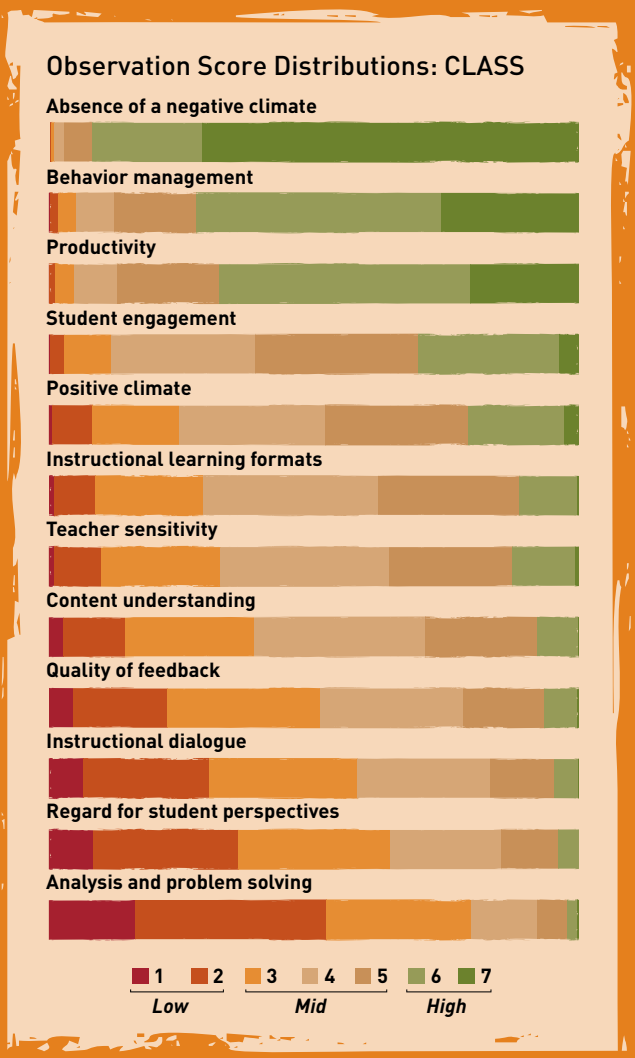
ACROSS THE INSTRUMENTS

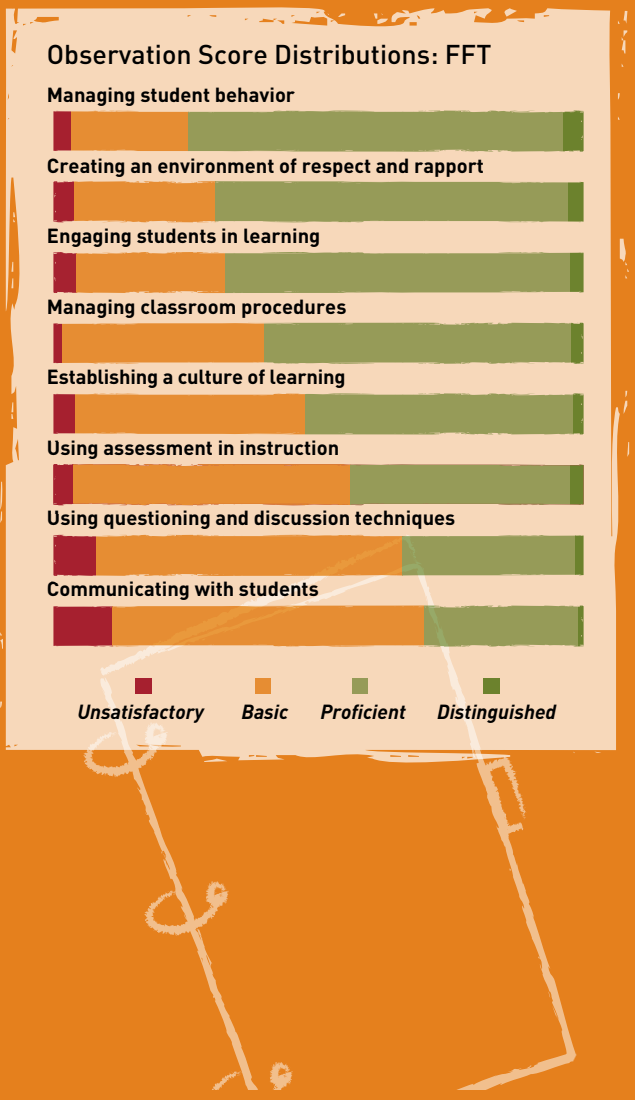
These classroom observation instruments, though they differ in emphasis, portray a remarkably consistent image of classroom practice. The MET project teachers tended to do fairly well at behavior management and time management. However, scores were lower in areas such as problem solving (CLASS), effective discussion (FFT), intellectual challenge (PLATO), richness (MQI), and investigation (UTOP). The task of developing conceptual understanding is complex, and these results suggest that as the teaching tasks grew in complexity, it grew rare.

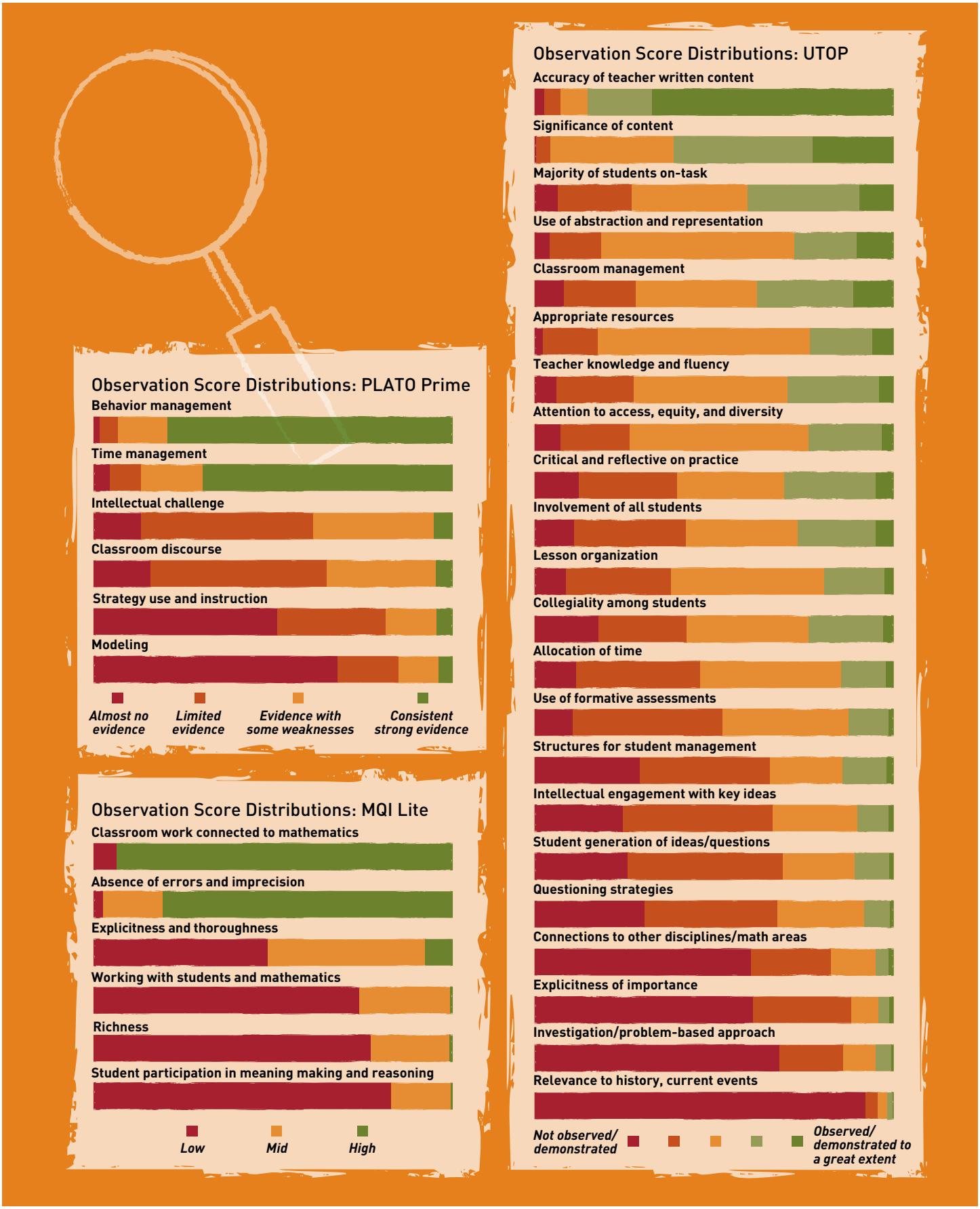
Figure 8. Observing Teaching Practice through Five Lenses

Each chart represents one of the five observation instruments used in the MET project to assess classroom practice. Each row represents the performance distribution for a particular competency. All of the instruments use scales with between three and seven performance levels. Higher numbers represent more accomplished performance. Each chart is organized by prevalence of accomplished practice, with the lowest ratings on the left (red) and the highest on the right (green). These data were drawn from ratings at the lesson or lesson-segment level, based on observing a total of 30 minutes of instruction (except for UTOP, for which raters observed more). For example, a rater using CLASS would give ratings for 15-minute lesson segments.

A few patterns are immediately visible in the data. First, raters judged the observed lessons to be orderly and generally on-topic. Across these instruments, behavioral-, time-, and materials-management competencies were rated as most accomplished. Second, across all instruments, raters rarely found highly accomplished practice for the competencies often associated with the intent to teach students higher-order thinking skills.






Observation Score Distributions: PLATO Prime
Behavior management
Time management
Intellectual challenge
Classroom discourse
Strategy use and instruction
Modeling
Almost no evidence
Limited evidence
Evidence with some weaknesses
Consistent strong evidence
Observation Score Distributions: MQI Lite
Classroom work connected to mathematics
Absence of errors and imprecision
Explicitness and thoroughness
Working with students and mathematics
Richness
Student participation in meaning making and reasoning
Low
Mid
High
Observation Score Distributions: UTOP
Accuracy of teacher written content
Significance of content
Majority of students on-task
Use of abstraction and representation
Classroom management
Appropriate resources
Teacher knowledge and fluency
Attention to access, equity, and diversity
Critical and reflective on practice
Involvement of all students
Lesson organization
Collegiality among students
Allocation of time
Use of formative assessments
Structures for student management
Intellectual engagement with key ideas
Student generation of ideas/questions
Questioning strategies
Connections to other disciplines/math areas
Explicitness of importance
Investigation/problem-based approach
Relevance to history, current events
Not observed/ demonstrated
Observed/ demonstrated to a great extent

The Principal Components of Instruction

Each observation instrument measures a different set of competencies. In the versions used in the MET project, UTOP identifies 22 competencies, CLASS identifies 12, FFT identifies eight, and MQI and PLATO each identify six. Each instrument is based on a particular theory of instruction, yielding differing hypotheses regarding the essential components of effective instruction. Theoretically, at least, a teacher could be strong on one competency and weak on another.

Each additional competency included in an instrument adds costs. Obviously, adding a competency requires training time and scoring time for observers. However, adding a competency also risks lowering the quality of data on all the other competencies if observers have already reached the limits of their ability to keep track. When observers are overtaxed by the cognitive load of tracking many different competencies at once, their powers of discernment could decline. (For an analogy, imagine the juggler who can keep four balls up in the air easily but who drops all the balls when a fifth is added.)

Therefore, choosing any instrument for classroom observation requires balancing two considerations: On one hand, when there are many competencies, feedback can be very fine-grained, and improvement efforts could be tuned to match a teacher's specific strengths and weaknesses. On the other hand, as the set of competencies grows longer, the quality of the feedback is likely to suffer, as observers are overtaxed. Occasionally, it may be useful to economize by combining or dropping competencies that commonly occur together, that prove to be too difficult to measure reliably, or that are unrelated to other outcomes.

NATURALLY OCCURRING CLUSTERS AND PERCEIVED CLUSTERS

When two competencies almost always occur together, as part of the same "package" of teaching practices, statisticians would say that these two competencies reflect a single "principal component" of instruction. When competencies are part of the same principal component, it is difficult to identify which is more important for student achievement. To untangle the independent effect of each competency from the other, one must observe teachers with a particular strength on one but not the other (and vice versa). Yet, when competencies are part of the same principal component, knowing their level on one competency tends to predict their levels on the others.

Table 9 reports the cross-tabulation of three competencies from the FFT instrument. The top panel cross-tabulates observer scores on two competencies that one might expect to occur together: *managing student behavior* and *managing classroom procedures.* (Both are in Domain 2, intended to capture aspects of the classroom environment.)

To assign the highest score on managing student behavior in FFT, an observer is expected to observe the following: "standards of conduct are clear to all students and appear to have been developed with student participation," "monitoring by teacher is subtle and preventive," and "teacher response to misbehavior is highly effective and sensitive to students' individual needs, or student behavior is entirely appropriate." Assigning the highest score on managing classroom procedures requires that observers see "groups of students working independently, are productively engaged at all times, with students assuming responsibility for productivity,"

"transitions are seamless, with students assuming some responsibility for efficient operation," and "routines for handling materials and supplies are seamless, with students assuming some responsibility for efficient operation."

Table 9. Cross-Tabulating Scores for Different Competencies in Framework for Teaching

Cross-Tabulating Scores on Two Competencies in the *Same* Domain

		MANAGING CLASSROOM PROCEDURES				
		1	2	3	4	TOTAL
Managing Student Behavior	1	143	101	19	0	263
	2	172	1051	435	0	1658
	3	29	920	4186	100	5235
	4	2	7	185	96	290
	TOTAL	346	2079	4825	196	7446
				Proportion with exact match:	0.74	
				Proportion with one point difference:	0.26	
				Proportion with 2+ point difference:	0.01	

Cross-Tabulating Scores on Two Competencies in *Different* Domains

		QUESTIONING AND DISCUSSION TECHNIQUES				
		1	2	3	4	TOTAL
Managing Student Behavior	1	119	137	7	0	263
	2	329	1110	213	6	1658
	3	391	3020	1777	47	5235
	4	16	91	155	28	290
	TOTAL	855	4358	2152	81	7446
				Proportion with exact match:	0.41	
				Proportion with one point difference:	0.52	
				Proportion with 2+ point difference:	0.07	

Note: The above cross-tabulations are drawn from the first rater's scores on each of the videos scored with FFT.

For 74 percent of the videos, observers gave the exact same score on these two competencies (the orange shaded region along the diagonal). For another 26 percent of the videos, observers gave a score that was one-point different on the instrument's four-point scale. For only 1 percent of the videos (the unshaded portion of the table in the top right and bottom left of the table), the rater scores were more than one point different. Therefore, to determine the relative importance of *managing classroom procedures* and *managing student behavior,* we have only this small number of videos for which there is a substantial difference in scores. But with small samples, there will be little statistical power, and results are often statistically insignificant.

The bottom panel reports the cross tabulation of two competencies that seem more distant from each other: *managing student behavior* and *questioning and discussion techniques.* Indeed, they come from different domains of the FFT instrument. As reported in the figure, there is less direct overlap in these two

competencies—with 41 percent of the videos receiving the same score on the two competencies, and roughly half the sample scoring one point different. Still, only 7 percent of the sample was given scores more than one point different on these two seemingly distinct skills.

The clustering of competencies could reflect a common underlying talent or personality trait or training; those with a strength (or weakness) in one area may tend to have a strength (or weakness) in another. But any naturally occurring clustering could also be exaggerated by observers, especially if they are struggling to keep track of many different competencies simultaneously. Suppose that an observer has learned, perhaps even subconsciously, that virtually all of those who scored well (or poorly) in questioning strategies also scored well (or poorly) in classroom management. When that happens, an observer may have a hard time noticing the few cases in which a teacher with good questioning skills showed weak classroom management. Once they form an impression of one competency, they may subconsciously infer that the other was also present, even if they did not see it explicitly.[21]

THREE PRINCIPAL COMPONENTS

When there are many different competencies to be measured, principal component analysis allows one to identify the ways in which individual competencies tend to cluster together. It tells us the number of independent "clusters," or linear combinations of competencies, that account for most of the variance in a given set of measures. In this study, there were three principal components—or three clusters of competencies for each instrument—that accounted for the lion's share of the teacher-level variation in scores on each of the instruments.[22]

First, for all the instruments, observers tended to rate individual teachers high (or low) relative to their peers on all the competencies simultaneously. We saw above that a teacher who excelled at *managing student behavior* also tended to receive high scores on a seemingly unrelated competency, *questioning techniques.* For each of the instruments, the first principal component was simply an equal-weighted average of all competencies. Indeed, this first component accounted for much of the variation in scores on all five instruments.[23]

Second, all the instruments contained a component reflecting competencies associated with classroom management and time management. Observers consistently noticed whether or not a classroom was well organized and orderly, and this competency frequently did not coincide with the other competencies observers were looking for. Using the scores from the CLASS instrument, for example, this second component depended heavily on two competencies: *behavior management* and *productivity.* In FFT, this second component depended heavily on *managing classroom procedures* and *managing student behavior.* Similar patterns were observed on all the instruments.

21 Something similar might occur if two competencies almost never occurred together. If those who are poor in their questioning strategies are almost always strong in their classroom management, an observer may see one competency and subconsciously conclude that the other was not present.

22 The proportion of the teacher-level variance in the individual competencies accounted for by the first three principal components was 82 percent for UTOP, 87 percent for MQI, 90 percent for CLASS, and 97 percent for FFT and PLATO.

23 The proportion of the teacher-level variance in the individual competencies accounted for by the first principal components was 46 percent for MQI, 59 percent for PLATO, 60 percent for UTOP, 73 percent for CLASS, and 91 percent for FFT.

Notably, we found a similar pattern when performing a principal component analysis on the Tripod student survey data. When aggregated at the teacher level, just three principal components accounted for most of the variance in student responses, with the first principal component capturing the teachers' overall performance averaged across all measures. Moreover, as with the classroom observations, the second component was strongly related to the items asking students about the level of control and respect a teacher attained in his or her classroom. For example, classrooms varied greatly in the proportion of students agreeing with statements such as, "We use time well in this class and we don't waste time" or "Students in this class treat the teacher with respect." Moreover, this was a distinct component, implying that the orderly classrooms did not always stand out on the other dimensions of teaching captured in the student survey.

In addition to the "overall" component and the component focusing on classroom/time management, each of the instruments contained a third component that was unique to that instrument. For example, the CLASS instrument, which places an emphasis on the emotional climate in a classroom, was able to discern differences between teachers in their ability to create a positive emotional climate and to be sensitive to students' emotional state. FFT, which has roots in a "constructivist" theory of instruction, emphasizes the importance of getting students to describe their own thinking in class, and raters tended to identify a third component related to teachers' questioning and assessment techniques. Raters using the UTOP instrument discerned a third cluster seemingly similar to the third cluster from FFT, including "student generation of ideas," "questioning strategies," "structures for student engagement," etc.

IS ONE DOMAIN MORE IMPORTANT THAN ANOTHER?

We have done some exploratory work, trying to identify the importance of specific competencies and domains to student achievement gains. For example, we tested whether there was enough evidence to reject the constraint that all domains within each instrument have the same relationship to student achievement as the others. However, we could only reject that hypothesis for one instrument in math (FFT) and one instrument in ELA classrooms (CLASS). For the remaining instruments, there was simply not enough evidence to reject the hypothesis that the incremental contribution of each domain to predicting student achievement gains was equal to all others. We will be pursuing this question further in future work. However, for the remainder of the analysis, we will look at the overall scores on each of the instruments.

CORRELATION AMONG INSTRUMENT SCORES

Table 10 reports the relationships among the overall scores on each of the five instruments. To distinguish the teacher-level correlation in the overall scores, as opposed to other sources of variation coming from rater error or particular lessons being observed, we report disattenuated correlations.[24] Disattenuated correlations

24 If a teacher is teaching in two different course sections, 1 and 2, and we have scores on two different instruments, X and Y, then we calculated the disattenuated correlation in scores as $\rho_{X,Y} = \frac{Cov(X_j, Y_{Notj})}{\sqrt{Cov(X_1, X_2)Cov(Y_1, Y_2)}}$. Therefore, the disattenuated correlations are based on the subset of teachers with scores from more than one course section.

correct for measurement error in any two measures being compared.[25] A disattenuated correlation equal to one implies that if one were to observe teachers many, many times on each of two instruments, the average score on each of the two instruments would be perfectly correlated. In other words, while any particular score on one of the instruments may be subject to rater error or variation from lesson to lesson, a disattenuated correlation equal to one means the instruments are seeking to measure two constructs that are either the same or will always come together. When averaged over many observations, teachers would be ranked exactly the same way. By contrast, if the disattenuated correlation is equal to zero, the instruments are measuring entirely unrelated constructs.

Table 10. Disattenuated Correlations Across Five Indices

	CROSS-SUBJECT		MATH-SPECIFIC		ELA-SPECIFIC
	CLASS	FFT	UTOP	MQI	PLATO
CLASS	1				
FFT	0.88	1			
UTOP	0.68	0.74	1		
MQI	0.69	0.67	0.85	1	
PLATO	0.86	0.93			1

Note: These correlations with UTOP and MQI are based on math classrooms, and correlations with PLATO are based on ELA classrooms. As indicated by the shaded cells, we were not able to calculate disattenuated correlations for single classroom teachers.

The correlation between the two general pedagogical instruments, FFT and CLASS, was 0.88, implying that those two instruments, if used many times on the same group of teachers, would provide very similar rankings among teachers. The set of competencies measured by the two instruments—even if they appear distinct— are very highly correlated.

Moreover, the correlation between the two math instruments, MQI and UTOP, was 0.85. Although the instruments need not yield identical scores on a given lesson, they would yield nearly identical rankings after many observations. Again, these two instruments are looking for highly correlated competencies.

These disattenuated correlations are based on the relationship between instrument scores in different course sections for the same teacher. As such, they are driven by consistent aspects of a teacher's practice across sections and not by idiosyncratic incidents within any given lesson. For example, if we had used scores of the same lesson for calculating these correlations, we might see scores correlated because raters are reacting to common video content: the same behavioral disturbance in class or an obvious math error by a teacher. However, these correlations imply that the teachers who scored unusually high in one course section or lesson using one instrument also tended to score unusually high in an entirely different section or lesson using a different instrument. The estimates in Table 10 imply that the persistent component of a teacher's practice is correlated, as opposed to raters' common reaction to a particular lesson or a particular incident (which could also be correlated).

25 The disattenuated correlation corrects for measurement error when comparing two measures. This is distinct from the "correlation with underlying value-added," which we discuss later in the report. The former corrects only for measurement error in the value-added measure.

However, the lower correlations between the general instruments and the subject-specific instruments implies that they are measuring somewhat distinct concepts. The four correlations between MQI and FFT or CLASS and between UTOP and FFT or CLASS, ranged from 0.67 to 0.74. This implies that the math-specific instruments were indeed looking for a set of competencies that were not as strongly correlated with the competencies in the general pedagogical instruments. In contrast, the disattenuated correlations between PLATO and FFT and PLATO and CLASS were 0.93 and 0.86, respectively, suggesting that the overall performance captured by this language arts instrument was more closely aligned with the competencies being measured in the general pedagogical instruments.

IMPLICATIONS

The field is at an early stage in the evolution of observation instruments. This is hardly surprising, since this is the first large-scale comparison of multiple instruments with the same group of teachers and their outcomes. Nevertheless, we hope these results will spark further improvements in instrument design and training.

There are a number of questions that remain unanswered.

- First, we need to determine whether the small number of principal components discerned by these instruments reflects a general fact about teaching—that is, teaching competencies do come in a small number of packages or clusters—or whether individual observers can only keep track of a few sets of competencies in their minds simultaneously.

If the former is true, greater parsimony in instrument design is warranted. If the latter is true, we need to find ways to more efficiently process the information from a classroom observation. For example, perhaps some of the things we ask observers to track in an observation—such as the level of student engagement—would be better measured by surveying students, so that observers can track other things. Alternatively, observers could specialize in tracking a subset of competencies and on different occasions observers could focus on different aspects of teaching.

- Second, we do not know which competencies are most susceptible to improvement and which description of the competencies meshes most powerfully with teachers' own understanding of the job.

For example, suppose there were two competencies that generally occurred together and that were equally related to student learning outcomes. But also suppose that teachers were more likely to improve on one than another given the appropriate feedback and training opportunities. All else being equal, we are better off measuring competencies that teachers are inspired to improve and can improve with the right supports.

- Third, as the new Common Core State Standards and assessments are introduced, we will need a new set of instruments more closely aligned with those standards.

For example, the new literacy standards require more attention to reading strategies and writing in subjects such as science and social studies. The new math standards require students to master a smaller number of core ideas in math. The next generation of observation instruments should more closely reflect those standards.

The Reliability of Classroom Observations

For observation results to provide adequate information on a teachers' skill set, we need to know that they reliably reflect consistent aspects of teaching practice and not the idiosyncrasies of a particular observer, lesson, or class. Most teachers and school leaders have in mind a system in which a single individual is watching a single lesson and providing feedback. In a system like that, the only determinant of reliability is "inter-rater reliability." The only question is: Would a different rater, watching the same lesson, on the same day, with this group of kids, come to the same judgment?

But while inter-rater reliability is important to this broader conception of reliability, it is not the only concern. Other things matter too, such as the *number* of different lessons to observe, or the number of different *sections* of students to observe a teacher with. If the goal is to present a reliable picture of a teacher's practice, these may matter just as much as the number of different raters doing the observing.

Focusing solely on inter-rater reliability ignores the possibility that the characteristics of a teacher's practice may vary from lesson to lesson, or from one group of students to another. Even if one had a high level of inter-rater reliability in scoring a given session, the system may still give a very unreliable assessment of a teacher's practice, if it does not demonstrate the full range of his or her skills in every lesson.

In addition, if inter-rater reliability becomes an end in itself, the incentive will be to focus on those aspects of practice that can be measured with a minimum of judgment: things such as, "Is the lesson objective written on the board?" "Did class start on time?" "Did class end on time?" or even "Was the teacher wearing a tie?" etc.

We analyzed results from a subset of the lessons that were scored by more than one rater. We studied the degree to which scores varied from teacher to teacher, section to section, lesson to lesson, and rater to rater. We did so in a way that allowed us to compare the extent to which different sources of variability affected overall reliability.

We decomposed the total variance in scores into each of its various components. **Table 11** reports the percentages of variance by factor for each of the instruments.[26] Below is a summary of our findings for each source of variance in the observation results:

- **Persistent Teacher Effects:** Across all of the instruments, 14 percent to 37 percent of the variance in scores is attributable to persistent differences among teachers. In other words, the vast majority of the total variance in scores across all lessons, teachers, and raters—typically around two-thirds—is due to factors *other than* persistent differences among teachers.[27]

26 These results are reported for those teachers with multiple sections, or classrooms, of students who they teach. As a result, they are drawn primarily from middle school classrooms and the subset of elementary teachers who specialize by subject.

27 The percentage attributable to teacher effects in MQI was lower, 7 percent to 16 percent. It may be that the MQI instrument was not well suited to the study design, with large numbers of raters provided online training and scoring using the panoramic video cameras. MQI raters noted problems with resolution of the board camera video (necessary for detecting errors in math instruction) about 20 percent of the time. However, different raters rarely flagged the same videos as problematic, so it is difficult to know if the problem lies with the video or the training or the instrument. ETS is currently evaluating the MQI instrument with another sample of math teachers.

Table 11. Decomposing the Variance in Instrument Scores

	PERCENTAGE OF VARIANCE					IMPLIED RELIABILITY		
	TEACHER	SECTION	LESSON	RATER	RESIDUAL	1 LESSON	2 LESSONS	4 LESSONS
UTOP (overall)	30	0	15	3	53	0.30	0.46	0.63
Class Environment	32	0	8	3	58	0.32	0.48	0.65
Lesson Structure	25	0	13	3	60	0.25	0.39	0.57
Implementation	24	0	12	2	62	0.24	0.39	0.56
Mathematics Content	16	0	18	6	60	0.16	0.28	0.44
CLASS (overall)	31	0	27	8	34	0.31	0.47	0.63
Emotional Support	28	0	28	10	34	0.28	0.44	0.61
Class Organization	32	3	18	14	32	0.32	0.47	0.62
Instructional Support	23	0	23	11	42	0.23	0.38	0.55
Framework for Teaching (overall)	37	4	10	6	43	0.37	0.53	0.67
Respect & Rapport	30	3	8	7	53	0.30	0.45	0.60
Questioning	15	4	12	6	62	0.15	0.25	0.38
Culture for Learning	25	0	10	7	57	0.25	0.40	0.58
Classroom Procedures	24	6	0	7	62	0.24	0.37	0.51
Communicating with Students	21	1	2	8	68	0.21	0.34	0.50
Managing Student Behavior	33	8	1	3	54	0.33	0.47	0.59
Engaging Students	20	3	12	6	59	0.20	0.33	0.47
Using Assessment	18	0	3	9	70	0.18	0.31	0.47
MQI (overall)	14	0	26	13	48	0.14	0.24	0.39
Connected to Math	12	0	43	2	43	0.12	0.21	0.35
Errors & Imprecision	6	0	5	16	74	0.06	0.11	0.20
Explicitness	16	0	0	21	63	0.16	0.27	0.42
Richness	8	0	24	10	58	0.08	0.15	0.26
Student Participation	13	0	27	11	49	0.13	0.23	0.37
Working With Students	7	0	25	13	55	0.07	0.13	0.23
PLATO (overall)	34	0	20	10	36	0.34	0.50	0.67
(overall standardized)	35	0	21	9	34	0.35	0.52	0.68
Intellectual Challenge	14	0	32	8	46	0.14	0.25	0.40
Classroom Discourse	18	0	28	8	45	0.18	0.31	0.47
Behavior Management	34	11	23	4	28	0.34	0.47	0.57
Modeling	13	0	15	14	57	0.13	0.24	0.38
Strategy Use & Instruction	14	0	5	20	61	0.14	0.24	0.39
Time Management	30	6	26	4	34	0.30	0.44	0.57

Note: The variance components above are reported at the video-level (as opposed to segment level). In the case of CLASS, FFT, and MQI, the same rater scored multiple segments per video. Therefore, for CLASS, FFT, and MQI, the video component above includes the video-level variance plus the segment-level variance divided by the number of segments. Similarly, the residual component includes the rater by video variance plus the segment-level residual variance divided by the number of segments. The reliability calculations were based on the assumption that a different rater was scoring each lesson (but the same rater was scoring all segments of a given video) and the lessons all came from the same course section.

- **Lessons:** For UTOP, PLATO, and FFT, the between-lesson variance in scores for teachers was roughly one-half as large as the variance among teachers.[28] For CLASS and MQI, the percentages were even higher. In other words, even if we had a very precise measure of the quality of instruction in *one lesson,* we would still have an inaccurate impression of a teacher's practice—because a teacher's score varies considerably from lesson to lesson. Therefore, to capture persistent differences in teacher practice, it is important to average across more than one lesson.[29]

- **Course Section:** Perhaps surprisingly, the course section (i.e., the identities of the students sitting in the classroom) played little role in the scores. In all five instruments, no more than 4 percent of the variance in the overall score was associated with course section.[30] When a teacher taught more than one section, two videos were drawn from one section and two videos were drawn from another section. The variation in scores between videos in the same section was nearly as large as the variation in scores between videos drawn from different sections.

- **Raters:** For most of the instruments, 10 percent or less of the total variance in scores was due to "main" rater effects—that is, when some raters consistently score high and other raters consistently score low. In other words, it seems possible to constrain tendencies to score too leniently or too harshly through training. The main exceptions were in MQI (particularly for *errors and imprecision* and *explicitness)* and to a lesser extent in PLATO (particularly for *modeling* and *strategy use and instruction)*. This may be due to the higher levels of subject knowledge required to score these competencies.

- **Residual:** For every instrument, the largest source of variance falls into the final category, what statisticians refer to as "the residual variance." Although the "main" rater effects are small (i.e., the mean scores across lessons vary little by rater), the large residual variance reflects the fact that raters often disagreed on any particular lesson (i.e., the rater-by-lesson and rater-by-segment variance are high). For example, on one lesson, rater A may give a higher score than rater B, and on another lesson, rater B may give a higher score than rater A.

ACHIEVING HIGH RELIABILITIES FOR OUR STUDY

Having determined the sources of variance for each instrument, we calculated the reliability of measurement at the teacher level when scores were averaged across different numbers of lessons and raters. We asked: "If we were to aggregate scores across different numbers of lessons and sections and observers, what portion of the variance in average scores would reflect consistent differences between teachers, rather than other factors such as raters or lessons?"

28 The percentages do not include the variation due to rater error; they are meant to approximate the variation one would observe even if one had a very precise score for every lesson.

29 Suppose we scored one lesson per teacher. Even if we scored that lesson an infinite number of times and drove the rater variance and residual variance to zero, the upper bound on teacher-level reliability would be $\frac{\%Teacher}{\%Teacher + \%Section + \%Lessons}$. For example, on the emotional support dimension of the CLASS instrument, the upper bound for teacher-level reliability based on a single lesson would be 0.55.

30 Less surprising, the highest section-level influences were observed for those competencies related to classroom management and student behavior. For instance, 8 percent of the variance in the managing student behavior competency in FFT was associated with course section.

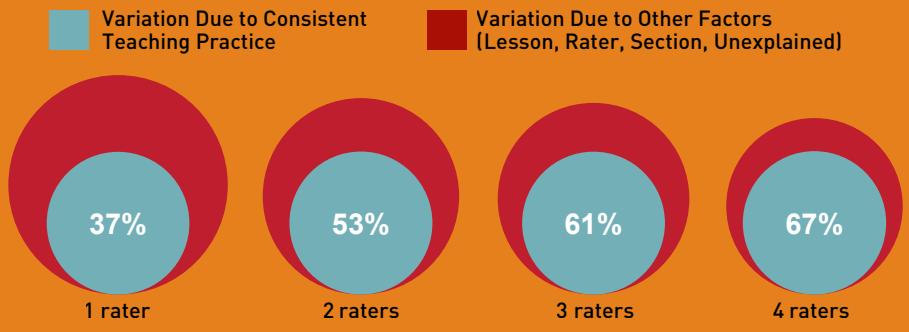
The two main obstacles to reliability proved to be (1) variation from lesson to lesson for a given teacher and (2) variation due to differing judgments of raters watching the same lesson. We therefore estimate the effect of using different numbers of lessons and raters. In the last three columns of Table 11, we estimate the reliability achieved with three different scenarios:

- Scoring one lesson by one observer;
- Scoring two lessons, each by a different observer; and
- Scoring four lessons, each by a different observer.[31]

On all the instruments we tested, a single observation by a single rater generated very low reliability: 0.37 or less. That means that just 37 percent of the total variance in scores from one observation reflects consistent differences in a teacher's practice. Unfortunately, there's no way to know whether any particular score is part of that 37 percent. In other words, a single rater watching a single lesson produced a very noisy assessment of a teacher's practice.

When we averaged scores from two different lessons, each scored by a different observer, our reliabilities increased. However, to achieve reliability in the neighborhood of 0.65 on the overall score for most instruments, we had to score four different lessons, each with a different rater. We were able to do so with UTOP, PLATO, CLASS, and FFT (although not MQI).[32] In **Figure 9,** results from FFT illustrate how adding additional observations led to higher reliability by decreasing variation due to other factors.

Figure 9. Multiple Observations Led to Higher Reliability



NOTES: The number in each circle is the percentage of variance in average FFT scores attributable to teacher effects. The area of the inner circle represents the variance in aspects of teachers' practice that is consistent across lessons, while the area of the outer circle adds in variation due to other factors, such as rater disagreement and lesson-to-lesson variance. As the number of observations increases, the variance due to consistent teaching practice remains constant, while the variance due to other factors declines, as it is averaged over more observations.

31 The reliability coefficients can be calculated using the following formula:

$$\frac{\%Teacher}{\%Teacher + \frac{\%section}{\#\,of\ sections} + \frac{\%lessons}{\#\,of\ lessons} + \frac{\%raters}{\#\,of\ raters} + \frac{\%residual}{total\ \#\,of\ scores\ per\ teacher}}$$

32 The outlier was the MQI instrument, for which we were only able to achieve a reliability of 0.39 with four lesson scores. This may indicate that the aspects of teaching emphasized by MQI are simply difficult to measure reliably.

IMPLICATIONS FOR PRACTITIONERS

We caution against extrapolating from our results too literally and concluding that a state or district would require such numbers to obtain similar reliabilities. The context of our study was unique in several obvious ways: (1) our observers were trained and required to demonstrate their ability to score accurately before they could begin scoring, (2) our observers had no personal relationship with the teachers being observed, (3) our observers were watching digital video and were not present in the classrooms, and (4) there were no stakes attached to the scores for teachers in our study (or for the observers).

Districts could discover that they require fewer than four (or, possibly, more than four) observations to achieve similar levels of reliability. On one hand, principals know much more about their teachers and their students than our observers did. Moreover, in-person observations could prove to be more reliable than our video because observers can see or hear the interactions in the classroom with more fidelity. For both reasons, real-world observations might bring to bear more information and, as a result, be more reliable. On the other hand, principals and peers may have a difficult time putting aside preconceived notions of a teacher's practice or may hesitate to be critical when there are stakes attached. (Recall that principals have rated most of the same individual teachers "satisfactory," often over many years.) If so, their observations could be even less reliable than ours. We cannot say which effect would predominate.

Nevertheless, our experience in scoring thousands of videos leads us to offer the following guidance to the many states and districts that are in the midst of redesigning the way they do classroom observations:

- First, to achieve acceptable levels of reliability with classroom observations, observers should demonstrate their ability to use the instruments reliably *before* observing.

Although good training is obviously necessary, it is unlikely to be sufficient. At the end of training, any prospective observer (principal, instructional coach, or peer teacher) should demonstrate that they can use an instrument fairly and reliably. This need not be a cumbersome and costly process. For instance, one way to do this would be to use a process similar to that used in the MET project, in which observers "pre-score" a set of videos using a particular observation instrument and then require observers to be able to reproduce those scores by watching those videos online. Another way would be to have a trained observer physically accompany each prospective observer on some initial observations and compare notes afterward. However this is done, fairness and reliability will require observers to demonstrate their understanding of the instrument at the end of training.

- Second, to produce reliable feedback on a teacher's practice, states and districts will need to observe a teacher during more than one lesson.

In our study, with trained raters, individual teachers' scores varied considerably from lesson to lesson. There are a number of possible reasons why this might be true. For example, different types of material may simply require teachers to showcase different skills; no one lesson provides a complete picture of their practice. The variation between lessons we observed was very unlikely to have been driven by any of the factors that made our study unique—the use of digital video, the use of trained observers, the use of impartial observers, or the absence of stakes. Rather, we expect this to be a general challenge that practitioners will also confront in their own observations.

- Third, to monitor the reliability of their classroom observations and ensure a fair process, districts and states will need to conduct some observations by impartial observers and compare the impartial scores with the original scores.

Most school districts will find it difficult to conduct impartial reviews for all their teachers. Fortunately, this is not necessary for tracking reliability. Simply having impartial raters provide scores for a representative subset of teachers (for instance, using a random "audit" for a small percentage of teachers) would be enough. A small sample of scores by impartial observers could be used as a quality control mechanism, a "dipstick" to measure the reliability of a system.

How a Reliability Audit Might Work

To check the reliability of official classroom observations, it is important to have at least a subset of observations done by impartial observers with no relationship to the teacher. Impartial observers are needed because they would not share any prior preconceptions another observer might have (positive or negative) due to past relationships. In addition, to shed light on the reliability across a district, the subset of teachers chosen for supplemental observation by impartial observers needs to be representative. Therefore, the two keys to checking system reliability are representative sampling and impartial observers.

Here's one way such a system might work: A district administrator could make a list of all teachers in all grades and subjects where observations are occurring and draw a random sample of, say, 100 teachers. The district could dispatch a set of observers from elsewhere in the system to perform one additional observation for each of the sampled teachers. These would not need to be done by "master observers"; the typical observer used in the system would be sufficient. And these additional observations need not count in any teacher's evaluations.

However, these observers should have no personal relationship with the teachers being observed and no prior exposure to the official observation scores. (They could be randomly assigned to the specific teacher they will observe as a further reassurance.) Also, the external observers would not need to be present on the same day that any of the official observations are being done. In fact, it would be better if they were not. The key is to get an independent assessment of the sample teachers' practice, using the same instrument used for the official evaluation, which is not subject to any of the other possible sources of variation that could influence the official observation.

Here's how a district or state could use such a process to produce an estimate of reliability of its system overall: Suppose E_j represents the single observation score for teacher j by the external, impartial rater. And suppose O_j is teacher j's official observation score. O_j could be the result of a single observation by a principal or the combination of multiple observations by multiple observers. The two most important requirements are, first, that they are measured using the same instrument with the same scale, and second, that any errors in measurement are plausibly independent. Then suppose T_j represents the underlying score for teacher j, his or her actual level of skill on the aspects of practice measured by the instrument. E_j and O_j are both imperfect measures of T_j, with errors ε_{1j} and ε_{2j} respectively. That is, $E_j = T_j + \varepsilon_{1j}$ and $O_j = T_j + \varepsilon_{2j}$. As long as the external observers are randomly assigned, did not consult with the official observer, and did not

observe on the same day as the official observer, then ε_{1j} and ε_{2j} should be independent of each other. (If the impartial observations are done once for the randomly selected teachers and the official scores are based on many observations, the two measures probably will have different error variances. This is fine, as long as the errors in measurement are independent.)

One could estimate the reliability in the official score with the following simple calculation, where $\bar{O}$ is the mean score for the official observation and $\bar{E}$ is the mean score for the extra observations:

$$Reliability\ of\ official\ scores = \frac{\sum_{j=1}^{100}(E_j - \bar{E})(O_j - \bar{O})}{\sum_{j=1}^{100}(O_j - \bar{O})^2}$$

Recall that reliability is the proportion of the variance in observation scores that reflects consistent differences between teachers and not differences between raters or lessons or course sections. The reason this calculation works is that the only shared source of variation in the supplemental observations and the official observations should be teachers' underlying skill level.[33] The covariance in the numerator provides an estimate of the variance in teachers' underlying skill—which is what we need to calculate reliability. Although districts would not have the perfect measure of effectiveness for any *individual* teacher, they could estimate the variance among teachers using the method above. (This same idea is used again in our discussion of underlying value-added in the next section.) The denominator in the reliability calculation is just the total variance in official scores. The ratio indicates the proportion of the variation in official scores that is due to the variance in teachers' actual skill level.

The sample size of 100 should suffice for estimating the reliability of the system *as a whole.* Generating finer estimates of reliability—by grade level or subject or even by school or principal—would require a larger number of additional observations.

33 Actually, if teachers are being observed in front of the same group of youth (and not in different course sections), any influence of a particular group of students on observer scores could be another source of shared variance, in addition to the teacher. As such, the suggested calculation could yield upwardly biased estimates of the reliability. If teachers are teaching in multiple sections, a school system might calculate the reliability only for observations done in different sections. However, our earlier finding that course section was a minor source of variation in video scores would lessen the bias.

Validating Classroom Observations with Student Achievement Gains

Underlying each of the instruments is a different vision of what effective teaching looks like. In this section, we test the alignment of each of those visions with measured student achievement gains. We look at gains rather than end-of-year scores because end-of-year tests scores partially reflect children's differing starting points. We want to test whether the *progress* students make is related to their teachers' instructional practice.

We measure gains in student achievement by comparing each student's end-of-year achievement with that of other students who had similar prior performance and demographic characteristics *and* who had fellow classmates with similar average prior performance and demographics. Often, researchers studying student achievement gains do not include additional controls for classroom peer characteristics. However, in our analysis, an individual student's growth was related to his or her average classmate's starting point. As a result, we control for mean baseline peer performance and other characteristics in this analysis.[34]

In our sample, the state assessments in math and ELA varied from district to district (since each of the districts was in a different state). To put the measures on a similar footing, we first standardized test scores to have a mean of zero and a standard deviation of one (for each district, subject, year, and grade level). We then estimated the following statistical model, controlling for each student's test score in that subject from the prior year, a set of student characteristics, and the mean prior test score and the mean student characteristics in the specific course section or class that the student attends:

$$S_{it} = X_{it}\beta + \overline{X}_{jkt}\gamma + \theta S_{it-1} + \lambda \overline{S}_{jkt-1} + \varepsilon_{it}$$

where the i subscript represents the student, j subscript represents the teacher, k subscript represents the particular course section, t subscript represents the year, X is a vector of student characteristics, $\overline{X}_{jkt}$ represents the mean of these student characteristics by class, S_{it-1} represents student baseline scores, and $\overline{S}_{jkt-1}$ represents mean student baseline scores in the class. We estimated separate specifications for each district, grade level, and subject (mathematics or ELA). The available student characteristics varied by district but included student demographics, free or reduced-price lunch status, English language learner (ELL) status, special education status, and gifted student status.[35]

To generate teacher-level value-added estimates ($\hat{\tau}^{S}_{jkt}$) for the test S, we first estimated the above specification and averaged the residuals by teacher, subject, and year (or teacher, section, subject, and year if a teacher taught more than one course section). In other words, the statistical model above produces an "expected" achievement for each student based on his or her starting point and the starting point of his or her peers in class. Some students "underperformed" relative to that expectation and some students "overperformed."

34 In the final report from this project, we will be able to test a number of alternative statistical models for estimating value-added using the data from 2009–10. We will then be able to identify the statistical model that is most accurate in predicting teacher impact following random assignment.

35 The student-level covariates used in the regressions included, in Charlotte-Mecklenburg, race, ELL status, age, gender, special education, gifted status; in Dallas, race, ELL, age, gender, special education, free or reduced-pricelunch; in Denver, race, age, ELL, free or reduced-price lunch, gender, and gifted status; in Hillsborough, race, ELL, age, special education, gifted status, and free or reduced-price lunch; in Memphis, race, ELL, free or reduced-price lunch, gender, gifted status, and special education; in NYC, race, ELL, gender, special education, and free or reduced-price lunch. Differences in covariates across districts may reduce the reliability of the value-added estimates.

In common usage, the term "achievement gain" is used to describe many different concepts, such as the change in a student's score from one year to the next. We use the term "achievement gain" in this report in a very specific way. A teacher's average student achievement gain is the average *difference between students' actual and expected achievement test score at the end of the year across all tested students in a classroom who have a prior year achievement test score.*[36] When we say that a teacher's students "moved ahead" or "fell behind"—we mean they overperformed or underperformed on end-of-year tests relative to students elsewhere in the district with similar characteristics and similar classroom peers at baseline.

As noted earlier in this report, in addition to state tests, students in participating classes took challenging supplemental performance assessments in spring 2010: Students in grades 4–8 math classes took BAM, while students in grades 4–8 ELA classes took the SAT9 OE reading assessment. In addition to being cognitively challenging, these tests were chosen because they have high levels of reliability and show evidence of fairness to members of different groups of students.[37] We generated teacher-level value-added estimates for the supplemental assessments using the same methods as for the state tests, with state test scores serving as the baseline regressors.

ARE THE MET PROJECT VOLUNTEERS REPRESENTATIVE OF TEACHERS IN THEIR DISTRICTS?

In **Table 12,** we report the mean student achievement gain estimates for the MET project teachers and the other teachers working in grades 4 through 8 in the MET project districts who did not volunteer for the MET project. The mean estimated student achievement gain in math and ELA was very similar for MET project and non-MET project teachers in 2009–10. On average, there's no evidence that the MET project teachers were achieving unusually high (or low) gains.[38] However, there is some evidence that the variance in student achievement gains was narrower among the MET project teachers: in math, the 25th and 75th percentiles were –0.136 and 0.135 (in student-level standard deviation units) for the MET project teachers and –0.165 and 0.164 for the other teachers in the districts; in ELA, the 25th and 75th percentiles were –0.090 and 0.101 for the MET project teachers and –0.138 and 0.127, respectively, for the other teachers.

36 This is very similar to a "random effects" specification for teachers, in which a teacher's ability to raise student achievement is assumed to be independent of the other covariates. In practice, since most of the variation in student baseline characteristics is observed within classrooms, rather than between classrooms, a model with random teacher effects yields estimates very similar to the fixed teacher effects model.

37 *Balanced Assessment in Mathematics (BAM):* Each of the test forms for BAM includes four to five tasks and requires 50–60 minutes to complete. Because of the small number of tasks on each test form, however, we were concerned about the content coverage in each teacher's classroom. As a result, we used three different forms of the BAM—the relevant grade level forms that were first released in 2003, 2004, and 2005—in each classroom. BAM uses question formats that are quite different from those in most state mathematics achievement tests. There is also some evidence that BAM is more instructionally sensitive to the effects of reform-oriented instruction than a more traditional test (ITBS). For sample items from the BAM test, see Appendix 1 in our previous research report *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.* Scores were standardized by test form and grade.

Stanford 9 Open-Ended (SAT9 OE) reading assessment: Each form of the SAT9 OE assessment consists of a narrative reading selection followed by nine questions. Students are required not only to answer the questions but also to explain their answers in writing. Sample items from the SAT9 OE exam are available in Appendix 2 of our *Initial Findings* research report.

38 Because we had to rely on volunteers, we were concerned that the MET project teachers may have included only exemplary teachers. The fact that they had student achievement outcomes similar to their colleagues' may reflect the general absence of performance measures available at the time they volunteered.

Table 12. Comparison of Value-Added of MET Project Volunteers with Other Teachers in Each District in 2009–10

		MEAN VALUE-ADDED ELA	ELA VALUE-ADDED (25TH,75TH)	MEAN VALUE-ADDED MATH	MATH VALUE-ADDED (25TH,75TH)
All Districts	MET	0.007	(-0.090, 0.101)	0.006	(-0.136, 0.135)
	Non-MET	-0.001	(-0.138, 0.127)	0.003	(-0.165, 0.164)
Charlotte Mecklenberg	MET	-0.003	(-0.083, 0.093)	0.013	(-0.128, 0.124)
	Non-MET	-0.011	(-0.111, 0.094)	-0.007	(-0.142, 0.133)
Dallas	MET	0.007	(-0.080, 0.097)	0.014	(-0.134, 0.146)
	Non-MET	-0.010	(-0.099, 0.106)	-0.002	(-0.162, 0.130)
Denver	MET	-0.012	(-0.132, 0.072)	-0.031	(-0.184, 0.131)
	Non-MET	0.008	(-0.098, 0.110)	0.008	(-0.130, 0.160)
Hillsborough	MET	0.005	(-0.061, 0.070)	-0.012	(-0.094, 0.073)
	Non-MET	-0.005	(-0.100, 0.096)	-0.005	(-0.104, 0.093)
Memphis	MET	0.020	(-0.138, 0.160)	0.017	(-0.193, 0.236)
	Non-MET	-0.018	(-0.142, 0.104)	0.005	(-0.223, 0.232)
New York City	MET	0.010	(-0.124, 0.128)	0.006	(-0.178, 0.165)
	Non-MET	0.003	(-0.156, 0.147)	0.006	(-0.181, 0.181)

Note: Table is based only on teachers of grades 4–8 in math or ELA. The subset of MET project teachers with scored videos had very similar means and distributions as all MET project teachers.

THE RELATIONSHIP BETWEEN STUDENT ACHIEVEMENT GAINS AND TEACHERS' UNDERLYING VALUE-ADDED

Of course, student achievement gains are an imperfect measure of a teacher's actual contribution to student learning. In any given year, student achievement gains may vary due to factors that have nothing to do with a teacher's practice. For example, any student could be sick (or sleepy or distracted) on the day of the test. Some of these factors are independent for each child and would average out with large numbers of students. However, especially in education, where a disruptive or unusually bright student could affect the learning of many peers, some factors would average out only over multiple course sections or multiple academic years. Prior research has suggested that both types of factors—the student specific and the classroom specific— can influence student achievement gains in any given year or course section.

In recent years, many have commented on the relatively low correlation in measured student achievement gains from year to year for a given group of teachers. For example, McCaffrey et al. (2009) used a five-year panel of teachers and students from five large school districts in Florida to study year-to-year correlations in student achievement gains in math. They found year-to-year correlations of 0.2 to 0.5. These are very similar to the year-to-year correlations in teachers' student achievement gains we observed among the MET project volunteers. The correlation in student achievement gains between 2008–09 and 2009–10 was 0.2 in ELA and 0.48 in math.[39]

39 Other estimates are reported in Harris and Sass (2006) and Koedel and Betts (2007).

Year-to-year correlations in this range are frequently cited as evidence that student achievement gains are "too volatile" or "too unreliable" to use for high-stakes decisions (Baker et al. 2010, Hill 2009). However, if student achievement gains are an imperfect measure of a teacher's underlying value-added in one year, they also will be subject to error in any other year. The year-to-year correlation is diminished by the fact that there is measurement error *in both years.*

What should interest us is the extent to which student achievement gains are related to teachers' underlying value-added. Yet this is *not* what the year-to-year correlations tell us: They tell us how one noisy measure is related to another noisy measure; not how either noisy measure is related to the truth.

Psychometricians (the statisticians who design tests) are familiar with this challenge. The tests they develop are imperfect measures of an underlying skill, not the skill itself. Suppose one has two imperfect measures, X_{j1} and X_{j2}, each measuring the same underlying skill, T_j, and each is subject to an independent (and similarly distributed) error in measurement, ε_{j1} and ε_{j2}: [40]

$$X_{j1} = T_j + \varepsilon_{j1}$$

$$X_{j2} = T_j + \varepsilon_{j2}$$

How are each of these imperfect measures, X_{j1} and X_{j2}, correlated to the underlying skill, T_j? Some simple algebra demonstrates that the correlation between either measure and the underlying score is the square root of the correlation between the two noisy measures.

To be sure, we never have a perfect measure of underlying value-added for any individual teacher. Nevertheless, we can calculate the correlation between an imperfect measure and that underlying concept simply by taking the square root of the year-to-year correlation (which, as long as the underlying skill is stable for short periods, also happens to equal the reliability).

Therefore, when it is reported that the correlation in a teacher's student achievement gains from one year to the next is 0.20 to 0.50, that is the same as saying that the correlation between the observed average student achievement gain in any year and a teacher's underlying value-added is a good deal higher—between 0.45 and 0.7. That correlation would be even stronger if the observed student achievement gain were averaged over several years.

To better understand how a measure's correlation to underlying skill is greater than the correlation between two years' results, consider an example from Major League Baseball. The correlation in players' batting averages from one year to the next is 0.58.[41] Like teachers, in any given year, a player's performance is likely to be affected by a number of random factors—weather, field conditions, quality of pitching, injuries—that have

40 These errors in measurement of underlying ability could reflect errors in the measurement of individual student's ability, sampling variation from year to year in the composition of the class, or any other nonpersistent factor affecting the measured performance of a whole class of students, such as "classroom chemistry," a bad flu season, or even a dog barking in the parking lot on the day of the test.

41 The year-to-year correlation in batting averages is even lower, 0.38, if one does not limit the sample to those players with at least 20 at bats in both seasons (author's calculations using data from 1871 to 2010). Modest year-to-year correlations in performance measurement are not isolated to teaching and baseball. As originally cited in Glazerman et al. (2010), a meta-study of objective performance measures in 22 highly complex occupations found year-to-year correlations in the range of 0.33 to 0.40 (Sturman et al. 2005).

nothing to do with his underlying skills. However, the correlation between single-year batting averages and *career* batting averages is considerably higher than the year-to-year correlation—approximately 0.75 (close to the square root of 0.58).

The analytic framework above is sometimes referred to as "true score theory." The noted psychometrician Melvin Novick wrote about it in 1966, although the concept was in use before then (Novick 1966). Throughout this report, we substitute the adjective "underlying" in place of "true" to acknowledge the fact that value-added is a measure of a specific underlying concept: persistent differences in measured student achievement gains associated with a teacher. It is possible that some portion of underlying value-added reflects unmeasured characteristics of the students persistently sorted into a teacher's class, not the causal effect of teachers. We will be testing this hypothesis in the final report. However, for the time being, it is important to recall that the measure of value-added we are using is based on nonexperimental data, using statistical controls to estimate teacher effects, and the term "underlying" is not meant as a synonym for "true."

Above, we discussed the case where two measures, X_{j1} and X_{j2}, are measuring a teacher's underlying value-added, T_j. That relationship can be generalized to the case where we have any proposed measure of effective teaching, Z_{ji}, (which could be student achievement gains with another group of students, or classroom observations, or a student survey, or a combination of all three). The correlation between any such measure and a teacher's underlying value-added, T_j, can be expressed as follows:

$$\rho_{z_{jt},T_j} = \frac{\sigma_{z_{jt},T_j}}{\sigma_{z_{jt}}\sigma_{T_j}} = \frac{\sigma_{z_{jt},T_j}}{\sigma_{z_{jt}}\sigma_{T_j}}\left(\frac{\sigma_{X_{jt}}}{\sigma_{T_j}}\right) = \frac{\rho_{z_{jt},x_j}}{\sqrt{\rho_{x_{j1},x_{jt}}}}$$

The final expression is just the correlation between the proposed measure (Z_{ij}) and student achievement gains, *divided by* the square root of the between-year correlation in achievement gains. In other words, the implied correlation of any measure of effective teaching with underlying value-added is equal to its correlation with student achievement gains relative to the correlation of underlying value-added with student achievement gains.

In this report, we will report two different ways to think about the correlation between the imperfect measure and the value-added measure: the correlation with observed student achievement gains ($\rho_{z_{jt}x_j}$) and the correlation with underlying value-added ($\rho_{z_{jt}T_j}$).

"POST-DICTING" A TEACHER'S MATH GAINS WITH THE STUDENTS IN THE PRIOR YEAR

When we validate each of the measures of instructional practice against student achievement gains, we ensure that the measures are drawn from *different groups of students.* Although we are able to include statistical controls for students' baseline test scores and demographics, there are many other factors for which we can't control but that may be having their own direct effect on measured achievement gains—such as unusual rambunctiousness on the part of some students at the time a particular lesson was recorded. We cannot control for those factors statistically, and yet an observer could see such factors when watching the videos. If such factors were reflected in observers' scores, then the relationship between measures of instructional practice and

student achievement gains could be exaggerated, specifically if that student trait were also positively related to student achievement gains. As a result, when we validate a measure of instructional practice, we are always doing so against a teacher's gains with a *different* group of students.[42]

Table 13 reports the relationships between a variety of measures collected during 2009–10 and teachers' student achievement gains in math during the prior year (2008–09). The first column reports the correlation of each measure with the prior year gain; the second column reports the implied correlation of the measure with underlying value-added.

Table 13. Relationship between Math Gains in Prior Year and Measures of Effective Teaching

MEASURES FROM 2010 USED TO PREDICT PRIOR YEAR GAINS	CORRELATION WITH PRIOR YEAR GAIN	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.18	0.25	0.08***
FFT	0.13	0.18	0.06***
UTOP	0.27	0.34	0.11***
MQI	0.09	0.12	0.05**
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.21	0.29	0.10***
FFT	0.18	0.26	0.09***
UTOP	0.34	0.43	0.16***
MQI	0.16	0.23	0.10***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.37	0.52	0.20***
FFT	0.36	0.50	0.16***
UTOP	0.46	0.59	0.21***
MQI	0.34	0.47	0.16***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The combined measures are equally weighted combinations of observation scores, student surveys, or value-added estimates. When combined, all components are standardized to have equal variance. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

The top panel reports the relationship for each of the observation instruments alone, based on each instrument's overall score. Teacher scores on CLASS based on four videos were correlated 0.18 with prior year gains and 0.25 with underlying value-added. The results were similar for the other cross-subject instrument, FFT, with correlations of 0.13 and 0.18, respectively. The correlations for one of the math-specific instruments,

42 While a worthwhile precaution, the use of different students for the instructional measures and outcomes is not foolproof. For instance, if a teacher is always assigned students with the same unmeasured trait, this strategy will not reduce the bias.

UTOP, was highest: 0.27 and 0.34, respectively.[43] The only measure with somewhat lower correlations was the MQI measure, which was correlated 0.09 with prior year gains and 0.12 with underlying-value-added. (This probably reflects the lower levels of reliability achieved with the MQI measure, noted in the previous section.)

The second panel *combines* each of the observation measures with the Tripod student survey results for each teacher. Whenever generating combined measures in this table, we weight each component equally. (We also standardize each component first by subtracting its mean and dividing by its standard deviation, so that the individual components have the same variance.)

Again, the results for CLASS and FFT were similar to each other, and UTOP seemed to show higher correlations: the unadjusted correlations with prior year gains were 0.21 and 0.18 for CLASS and FFT, respectively, and 0.34 for UTOP; the correlations with underlying value-added were 0.29 and 0.26 for CLASS and FFT, respectively, and 0.43 for UTOP. The relationships were again weakest for MQI, with an unadjusted correlation of 0.16 and a correlation with underlying value-added of 0.23.

The bottom panel reports the correlations achieved when combining all three types of measures: observation scores, Tripod, and math gains. In such cases, the correlations achieved with math gains were clustered around 0.4, with implied correlations with underlying value-added of 0.47 to 0.59.

The last column in Table 13 reports the difference in achievement gains in 2008–09 for those identified in the top and bottom quartiles based on the 2009–10 indicators of effective teaching. With the observation instruments alone, we are able to identify statistically significant, but modest, differences in student achievement gains in 2008–09: 0.05 to 0.11 student-level standard deviations. Adding the Tripod data improved the predictive power somewhat: the difference between the top and bottom quartiles expanded to between 0.09 and 0.16 standard deviations. However, we were able to identify larger differences using the three-measure combination of value-added, student feedback, and observations. By combining these, the difference in student achievement gains in another school year for those identified in the top and bottom quartiles was 0.16 to 0.21 standard deviations. The largest of these is nearly equivalent to the effect of a full year of schooling (0.25 standard deviations).

PREDICTING MATH GAINS IN ANOTHER SECTION

Table 14 reports results using indicators of effective teaching from one course section and comparing against outcomes from a second course section (both taught by the same teacher in 2009–10). These results draw primarily from those teaching in middle school grades, where the vast majority of teachers work with more than one group of students during a year (although a subset of the elementary teachers also specialized by subject and taught more than one section of students).

43 The UTOP results are not directly comparable to the results for the other instruments because they are based on different samples of teachers. We only have UTOP scores on the subsample of 250 teachers included in the Plan B sample (described in Appendix Table 1). Puzzlingly, MQI scores for Plan B teachers were more correlated with student achievement gains in 2008–09 than we report in Table 13 for the larger sample. However, the correlation to student achievement gains in another course section (reported in Table 14) and reliability (reported in Table 11) were generally not sensitive to the difference in sample for MQI or the other instruments.

Table 14. Relationship between Math Gains in Other Classroom and Measures of Effective Teaching

MEASURES FROM 2010 USED TO PREDICT GAINS IN ANOTHER CLASSROOM	CORRELATION WITH OTHER CLASSROOM GAIN	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.17	0.24	0.10***
FFT	0.13	0.19	0.07***
UTOP	0.18	0.26	0.07**
MQI	0.11	0.16	0.05**
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.28	0.40	0.14***
FFT	0.26	0.37	0.13***
UTOP	0.32	0.45	0.16***
MQI	0.25	0.35	0.12***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.43	0.61	0.20***
FFT	0.41	0.59	0.21***
UTOP	0.42	0.59	0.20***
MQI	0.40	0.57	0.19***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The combined measures are equally weighted combinations of observation scores, student surveys, or value-added estimates. When combined, all components are standardized to have equal variance. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

In most cases, the results were very similar to those obtained using prior year gains. When using the observation scores alone, the correlations with gains in the other section ranged from 0.11 to 0.18, implying correlations with underlying value-added of 0.16 to 0.26. Adding Tripod scores increased the correlations with gains to 0.25–0.32 and with underlying value-added to 0.35–0.45. The combination of three measures—observation instrument, Tripod scores, and student achievement gains in another section—increased the correlation with gains to 0.40–0.43 and with underlying value-added to 0.57–0.61.

OTHER OUTCOMES

Scoring well on state tests is not the only possible goal of public education. As a result, we used the same measures from above and asked, "How well did teachers' students do on other outcomes?"

As mentioned earlier, we supplemented the state tests in each of the districts with BAM, a test explicitly designed to assess conceptual understanding in mathematics and that uses items with very different formats than most state tests. We calculated student achievement gains on the BAM test using the same methodology we used for the state tests, that is, using the state test as the baseline control. We then calculated the mean student achievement gain of teachers in each quartile based on each of the measures of effective teaching reported earlier.

The first column of **Table 15** reports the differences in BAM gains for those in the top and bottom quartiles on each of the observation instruments. The difference in BAM gains were all statistically significant and positive, ranging from 0.05 to 0.11. When combining observations, Tripod scores, and value-added on the state tests, the difference in BAM scores was somewhat larger, 0.12 to 0.13.

Table 15. Differences in Other Student Outcomes in Math Classrooms

MEASURES FROM 2010 USED TO PREDICT OUTCOMES IN ANOTHER CLASSROOM	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES IN:		
	BAM GAIN	STUDENT EFFORT	POSITIVE EMOTIONAL ATTACHMENT
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.05*	0.07**	0.16***
FFT	0.08***	0.11***	0.16***
UTOP	0.11***	0.13***	0.20***
MQI	0.08***	0.04***	0.03***
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.11***	0.24***	0.45***
FFT	0.09***	0.25***	0.45***
UTOP	0.15***	0.23***	0.50***
MQI	0.08**	0.21***	0.41***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.12***	0.23***	0.44***
FFT	0.13***	0.24***	0.46***
UTOP	0.13***	0.23***	0.51***
MQI	0.12***	0.22***	0.39***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The combined measures above reflect equally weighted combinations of observation scores, student surveys, or value-added estimates. All three components are standardized to have equal variance. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units.

Beyond test scores, many parents care about whether their children are developing a positive emotional attachment to school. Children who enjoy school, they hope, are more likely to continue their education. In addition, parents want their children to be engaged in their learning. One measure of a student's engagement is the level of effort he or she puts in. We wouldn't want to emphasize practices that improve achievement but that make students miserable to be in school or to work less hard.

Therefore, we identified items in the Tripod survey that asked students to report their level of agreement with the following questions:

Effort

- "My teacher pushes us to think hard about things we read."
- "When doing schoolwork for this class, I try to learn as much as I can and I don't worry how long it takes."
- "I have pushed myself hard to understand my lessons in this class."

Emotional Response to School

- "This class is a happy place for me to be."
- Or the opposite of "Being in this class makes me feel sad or angry."

To be consistent with the test-based student outcomes, we generated "value-added" versions of the student effort and emotional response variables. Specifically, we constructed an index for each, weighting responses from "strongly disagree" to "strongly agree" with a numerical value of one through five. We standardized both indices to have a mean of zero and standard deviation of one at the individual student level. We then controlled for baseline achievement on the state tests, demographics, and classroom peer characteristics, using the same methods described for the state tests above.[44] In other words, we have adjusted the student effort and emotional response outcomes to reflect the fact that students with higher baseline achievement or other demographic characteristics may be more likely to report high levels of effort or positive emotional attachment.

As reported in the second and third column of Table 15, the difference in the top and bottom quartile teachers was also quite large on these outcomes. After combining classroom observations, Tripod scores, and value-added from one classroom and then arranging teachers into quartiles, the differences in student effort in other classrooms taught by top and bottom quartile teachers were 0.22 to 0.24 student-level standard deviations. The students were 0.39 to 0.51 standard deviations more likely to report a positive emotional association with being in the teacher's class.

CRITERION-BASED WEIGHTING VERSUS EQUAL WEIGHTING

When combining value-added, classroom observations, and student feedback in Tables 13 to 15, we weighted each measure equally: When we included two of the three, the weights were 0.50/0.50; when combining all three, the weights were 0.33/0.33/0.33. As we reported, regardless of the measure used, all the combinations were associated with student achievement gains.

States and districts must decide how to weight the various components of an evaluation. This has been the subject of considerable discussion in some states. We will explore that issue more deeply in our next report. However, to illustrate some of the trade-offs involved in that decision, we generated a set of criterion-based weights. Specifically, we first regressed each teacher's value-added on the state test from one course section against the three different measures from another course section: observation scores, student feedback, and value-added on the state test. This implicitly mimics a question any supervisor might ask: "What does a teacher's complete track record—on state tests, classroom observations, and student feedback—say about their likelihood of seeing large student achievement gains with *future* students?"

We can do something similar: We use data from one section of students to predict outcomes for *another* group of students in 2009–10. In so doing, we use the regression to determine the weights that would have done the best job in predicting.

The top panel of **Table 16** reports the predictive power and reliability of each individual measure taken alone. Each measure has a different combination of strengths and weaknesses:

44 The one difference was that we included prior scores in both math and ELA in the set of control variables.

- Not surprisingly, the value-added measure has the highest correlation with underlying value-added. However, it has lower reliability than student perceptions.
- The student perception measure, by itself, has the highest reliability, but it is not as correlated with underlying value-added as the value-added measure itself.
- The observation score measures generally have moderate reliability and modest correlations with value-added. Nevertheless, they provide diagnostic feedback.

Table 16. Comparing Use of Equal Weights and Criterion-Based Weights to Form Combined Measures in Math

	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES IN:				
		MATH GAIN	BAM GAIN	STUDENT EFFORT	POSITIVE EMOTIONAL ATTACHMENT	RELIABILITY
USING EACH MEASURE ON ITS OWN:						
Student achievement gains	0.69	0.24***	0.11***	0.14***	0.19***	0.48
Tripod survey	0.37	0.13***	0.06*	0.27***	0.57***	0.65
CLASS	0.24	0.10***	0.05*	0.07**	0.16***	0.43
FFT	0.19	0.07***	0.08***	0.11***	0.16***	0.40
UTOP	0.26	0.07**	0.11***	0.13***	0.20***	0.42
MQI	0.16	0.05**	0.08***	0.04***	0.03***	0.20
COMBINING OBSERVATIONS WITH GAINS AND TRIPOD: *Equal weights*						
CLASS	0.61	0.20***	0.12***	0.23***	0.44***	0.67
FFT	0.59	0.21***	0.13***	0.24***	0.46***	0.64
UTOP	0.59	0.20***	0.13***	0.23***	0.51***	0.65
MQI	0.57	0.19***	0.12***	0.22***	0.39***	0.55
Criterion-based weights						
CLASS	0.72	0.24***	0.12***	0.22***	0.36***	0.58
FFT	0.72	0.25***	0.13***	0.22***	0.37***	0.57
UTOP	0.66	0.22***	0.08*	0.24***	0.51***	0.61
MQI	0.71	0.24***	0.13***	0.22***	0.36***	0.56

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. Criterion-based weights are based on regressing value-added in state tests for a given teacher from one course section on each of the measures from another course section. The difference in mean gains between top and bottom quartile classrooms is reported in student-level standard deviation units. For achievement test gains, a difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

In the next panel in Table 16, we compare the equally weighted combinations against the criterion-weighted alternatives. We do so for all four instruments that measure math. The criterion-based weights placed higher weight on value-added on the state test and lesser weights on student feedback and observations. For example, when using CLASS as the classroom observation instrument, the weights on value-added, observations, and Tripod results were 0.729, 0.092, and 0.179, respectively, rather than 0.33, 0.33, 0.33. The coefficients for the other instruments are reported in Appendix Table 2.

As reported in Table 16, the correlations with underlying value-added are about 20 percent higher when using the criterion-based weights. (These are higher since the weights have been chosen to maximize their predictive power with respect to value-added.) For example, the correlation between the combined measure and underlying value-added increases from 0.61 to 0.72 when using CLASS. The differences in value-added between the top and bottom quartiles also rises about 20 percent. (These differences between top and bottom quartiles are roughly proportional to the estimated correlation with underlying value-added.) In other words, when student achievement is the criterion, a criterion-weighted combination of measures will predict better than an equally weighted measure. (Beyond the equal weighting and criterion-based weighting strategies discussed here, we will be releasing a report on alternative approaches to weighting in mid-2012.)

Interestingly, there is very little change in the correlation with BAM scores when moving from the equally weighted measure to the criterion-weighted measure (even when using state test scores as the criterion). Apparently, a teacher's gains on the state tests are about as predictive of gains on BAM as the combination of Tripod and observations—so that the change in weights did not diminish predictive power. The one exception was with UTOP, which seemed to be more strongly related to BAM gains than the other instruments. When the criterion-based weighting down-weighted the UTOP score, the difference in BAM scores between top and bottom performers fell. But UTOP was the only instrument where that was the case.

However, there are trade-offs to be considered. For example, the criterion-weighted measure is somewhat less correlated with student effort. In the combined measure incorporating FFT, the difference in average student effort in the top and bottom quartiles falls from 0.24 to 0.22. The difference in positive emotional attachment falls from 0.46 to 0.37. Both results derive from the fact that the criterion-based measure places less weight on the student feedback survey; because the outcomes are drawn from the student survey (albeit from a different course section), the Tripod student feedback is more correlated with the student-reported outcomes.

Important to note, the single-section reliability of the combined measure falls from 0.64 to 0.57 when moving from the equally weighted measure to the criterion-based measure.[45] The primary reason for this is that the student survey data are more consistent (and therefore more reliable) across sections than value-added or classroom observations. Because the criterion-based measure puts less weight on student feedback than the equally weighted version, reliability declines.

RESULTS FOR ENGLISH LANGUAGE ARTS

In Tables 17 through 20, we report analogous results using student achievement gains on the state ELA tests. However, in the ELA classrooms, we used PLATO as the subject-specific instrument, rather than MQI or UTOP.

Qualitatively, the results are similar. However, quantitatively, all of the results are somewhat weaker. In **Table 17,** we replicate the use of equally weighted combinations of 2009–10 data and compare those to student achievement gains on the state ELA tests in 2008–09. When using the observations alone, the correlation with underlying value-added is 0.12, 0.11, and 0.9 for CLASS, FFT, and PLATO, respectively. Adding the Tripod measure boosts these correlations to 0.21–0.23. Adding in value-added from 2009–10 raises the correlations further to 0.45–0.46.

45 The reliability when combining across multiple sections in a given year would be higher for both.

Table 17. Relationship between ELA Gains in Prior Year and Measures of Effective Teaching

MEASURES FROM 2010 USED TO PREDICT PRIOR YEAR GAINS	CORRELATION WITH PRIOR YEAR GAIN	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.08	0.12	0.03**
FFT	0.07	0.11	0.03**
PLATO	0.06	0.09	0.01***
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.14	0.21	0.06***
FFT	0.14	0.22	0.06***
PLATO	0.15	0.23	0.07***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.29	0.45	0.12***
FFT	0.30	0.46	0.12***
PLATO	0.30	0.46	0.11***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

In **Table 18,** we report analogous results, using student achievement gains on the state ELA test in another course section as the outcome. The results are very similar to Table 17. The main difference is that the subject-specific instrument, PLATO, seems to perform better than the cross-subject instruments, FFT and CLASS, in terms of its association with student achievement gains in another section. Taken alone, the correlation with underlying value-added is 0.10 and 0.11 for CLASS and FFT, respectively, and 0.24 for PLATO. The equally weighted combination of student achievement gains, Tripod, and observations were correlated with underlying value-added at 0.38 and 0.40 for CLASS and FFT, respectively, and 0.52 for PLATO. (We do not have a clear explanation of why PLATO performs better when the outcome is "other section" than when the outcome is "prior year.")

In **Table 19,** we report the differences between the top and bottom quartile teachers on the equally weighted measures using three alternative outcomes: student achievement gains on the SAT9 OE assessment, student-reported effort, and student-reported positive emotional attachment. Whether classroom observations are taken alone, or combined with student feedback and student achievement gains in ELA, the top classrooms on the combined measures had better outcomes on all three alternative outcomes than those in the bottom quartile of classrooms.

In **Table 20,** we replicate the comparison of equally weighted combinations (in the top panel) against criterion-based combinations (in the bottom panel). The results are similar to those we observed in mathematics:

- Taken alone, the individual measures each have strengths and weaknesses. The value-added measure is most highly correlated with underlying value-added, but it is also the least reliable. The student perception measure is the most reliable but not the most correlated with underlying value-added. The observation instruments, by themselves, fall between these two extremes.

- The criterion-based measures have correlations 22 percent higher on average with underlying value-added on the state tests than the equally weighted measures.
- The choice between criterion-based weighting and equal weighting makes little difference for identifying teachers with high value-added on the supplemental assessment, the SAT9 OE.
- When optimized to predict value-added on state ELA assessments, the criterion-based combination was somewhat less related to student-reported effort and positive emotional attachment to class than the equally weighted combination.
- Largely because the criterion-based weights put less emphasis on the Tripod data, the criterion-based combination was less reliable than the equally weighted measure. For example, the single-section reliability of the combination, including FFT, fell from 0.52 to 0.34.

Table 18. Relationship between ELA Value-Added in Other Classroom and Measures of Effective Teaching

MEASURES FROM 2010 USED TO PREDICT GAINS IN ANOTHER CLASSROOM	CORRELATION WITH OTHER CLASSROOM GAIN	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.04	0.10	0.01
FFT	0.05	0.11	0.02***
PLATO	0.11	0.24	0.04**
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.09	0.21	0.04**
FFT	0.11	0.23	0.05***
PLATO	0.16	0.35	0.05***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.17	0.38	0.07***
FFT	0.18	0.40	0.07***
PLATO	0.23	0.52	0.10***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

Table 19. Differences in Other Student Outcomes in ELA Classrooms

MEASURES FROM 2010 USED TO PREDICT OUTCOMES IN ANOTHER CLASSROOM	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES IN:		
	SAT9 GAIN	STUDENT EFFORT	POSITIVE EMOTIONAL ATTACHMENT
OBSERVATION INSTRUMENTS ALONE			
CLASS	0.10***	0.09***	0.18***
FFT	0.14***	0.12***	0.18***
PLATO	0.16***	0.08***	0.05***
OBSERVATION INSTRUMENTS WITH TRIPOD			
CLASS	0.09**	0.22***	0.43***
FFT	0.11***	0.22***	0.40***
PLATO	0.12***	0.21***	0.37***
OBSERVATION INSTRUMENTS WITH TRIPOD AND GAINS			
CLASS	0.10**	0.18***	0.35***
FFT	0.13***	0.22***	0.33***
PLATO	0.15***	0.20***	0.34***

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. The difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units. A difference of 0.25 standard deviations is approximately equivalent to one year of schooling.

Table 20. Comparing Use of Equal Weights and Criterion-Based Weights to Form Combined Measures in ELA

	CORRELATION WITH UNDERLYING VALUE-ADDED	DIFFERENCE BETWEEN TOP AND BOTTOM QUARTILES IN:				
		ELA GAIN	SAT9 GAIN	STUDENT EFFORT	POSITIVE EMOTIONAL ATTACHMENT	RELIABILITY
USING EACH MEASURE ON ITS OWN:						
Student achievement gains	0.45	0.08***	0.09**	0.07**	0.09*	0.20
Tripod survey	0.25	0.05***	0.05**	0.25***	0.51***	0.65
CLASS	0.10	0.01***	0.10***	0.09***	0.18***	0.43
FFT	0.11	0.02***	0.14***	0.12***	0.18***	0.40
PLATO	0.20	0.04**	0.16***	0.08***	0.05***	0.38
COMBINING OBSERVATIONS WITH GAINS AND TRIPOD: *Equal weights*						
CLASS	0.38	0.07***	0.10**	0.18***	0.35***	0.51
FFT	0.40	0.07***	0.13***	0.22***	0.33***	0.52
PLATO	0.45	0.08***	0.15***	0.20***	0.34***	0.5
Criterion-based weights						
CLASS	0.51	0.09***	0.11***	0.18***	0.30***	0.35
FFT	0.49	0.08***	0.11***	0.16***	0.28***	0.34
PLATO	0.51	0.09***	0.13***	0.16***	0.25***	0.36

Note: A *,**, or *** indicate a difference that is significantly different from zero at the 0.10, 0.05, and 0.01 levels, respectively. Criterion-based weights are based on regressing value-added in state test from one class on measures from the other class. For achievement gains, the difference in mean gains between top and bottom quartile classrooms are reported in student-level standard deviation units.

Risks of Mislabeling

Every personnel decision holds consequences, and yet every measure that might inform them is imperfect. For example, under most collective bargaining agreements, school administrators must decide which teachers to promote to tenure. There is no way to avoid deciding. A nondecision is effectively a decision to grant tenure, since that is usually the default outcome. Yet there are risks: Granting tenure to an ineffective teacher negatively affects student learning. Likewise, denying tenure to an effective teacher would be unfair to the teacher and costly to future students.

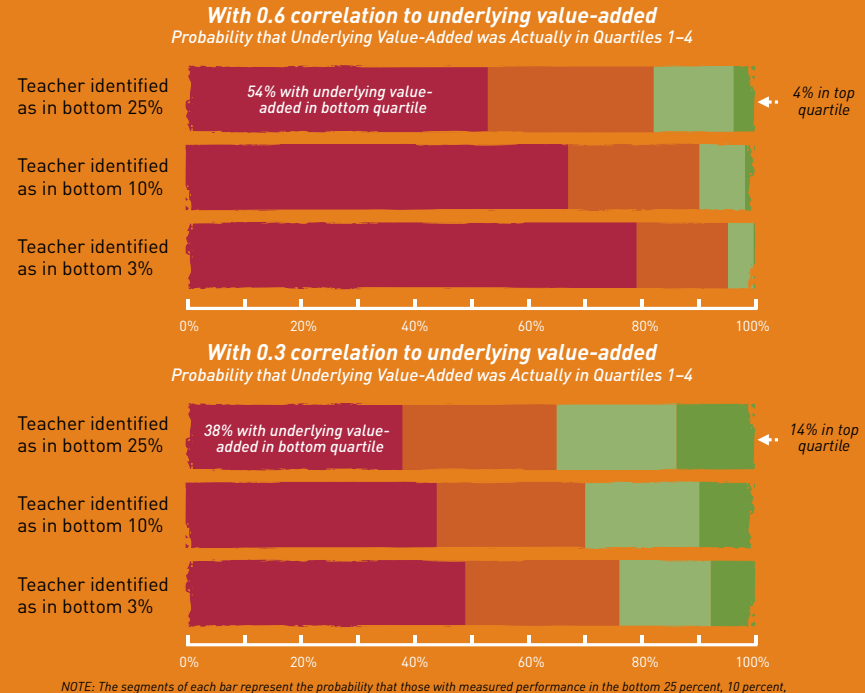
While we can't avoid this quandary, we can reduce the risk of error with better information. By estimating the extent to which a measure correlates with teachers' underlying value-added we can also estimate the likelihood of identifying teachers with high and low underlying value-added.

Take, for example, the combined measure we analyzed in our study that includes student feedback, an average of results from observations, and student achievement gains on state tests. By comparing a teacher's results on the combined measure from working with one group of students to the same teacher's student achievement gains from working with another group of students, we estimated a correlation with underlying value-added of 0.6 (with 1 representing a perfect correlation, and 0 no correlation).

As reported in **Figure 10,** when the correlation with underlying value-added is 0.60, 54 percent of the teachers in the bottom quartile in measured performance will come from the bottom quartile of underlying value-added. This is more than double the percentage that would have occurred by chance (25 percent), but it is not certainty. However, 82 percent of the bottom quartile teachers will have been drawn from the bottom half of underlying value-added; only 18 percent from the top half; and only 4 percent from the top quarter.

Of course, the bottom 25th percentile may not represent a realistic threshold at which consequential decisions are made, particularly if the vast majority of teachers perform in the middle ranges. Therefore we ask a different question: "What is the chance that a teacher in the bottom 10 percent or bottom 3 percent is truly in the bottom quarter or top quarter in terms of underlying value-added?" With a correlation with underlying value-added of 0.6, 67 percent and 79 percent of the bottom 10 and bottom 3 percent, respectively, will have underlying value-added in the bottom quarter; 2 percent and 0.5 percent of these will have underlying value-added in the top quarter.

Figure 10. Risk of Misclassification



NOTE: The segments of each bar represent the probability that those with measured performance in the bottom 25 percent, 10 percent, and 3 percent had underlying value-added in each quartile. Probabilities were simulated under the assumption that underlying value-added and measurement error were normally distributed.

To demonstrate why the correlation to underlying value-added is important to minimizing the risk of error, we repeat the exercise when the correlation with underlying value-added is 0.3. When the correlation is lower, the risk of making a mistake increases. Out of the bottom 10 percent in measured performance, 44 percent will have underlying value-added in the bottom quartile and 10 percent will have underlying value-added in the top quarter.

Therefore, as the correlation with underlying value-added increases, the likelihood of making mistakes in identifying teachers with high and low underlying value-added declines. However, the probability of making a mistake is never zero.

IMPLICATIONS

1. **Each of the observation instruments is related to student achievement gains, but the correlations were modest when observations were used alone.**

All of the observation instruments we tested were related to student achievement gains in math or ELA. However, the correlations are modest in size. This has two implications: First, evaluating based on observations alone will provide little leverage for raising student achievement. The difference in average student achievement gains for those in the top and bottom 25 percent on the observation instruments reported in Tables 14 and 18 was only 0.05–0.10 standard deviations for math and 0.01–0.04 for ELA. Second, in terms of underlying value-added, we run a higher risk of making a mistake, for example either promoting an ineffective teacher or failing to promote an effective teacher, if we do so based on observations alone.

2. **In some cases, the subject-specific instruments outperformed the general pedagogical instruments, but the results are mixed at this point.**

We found some evidence suggesting that the UTOP instrument was more correlated with underlying value-added in mathematics than the general pedagogical instruments, FFT and CLASS. However, the other math-specific instrument, MQI, was less correlated with student achievement gains in math than the general pedagogical instruments. In ELA, the PLATO instrument was more correlated with gains in another classroom in 2009–10 but was not more correlated with prior year student achievement gains than FFT or CLASS.

Currently, we see little reason to choose different classroom observation instruments in math and ELA. The general instruments, designed to be used in a wide range of grades and subjects, seem just as correlated to student achievement gains in math and ELA as the subject-specific instruments overall. This could change as the instruments are revised in coming years.

3. **The combination of observations and student feedback was moderately correlated with student achievement gains in math and ELA, which may allow for coverage in the nontested grades and subjects.**

We found that combining any of the observation instruments with the Tripod student survey data improved correlations with student achievement gains. Because we did not attempt to measure student achievement gains in any of the traditionally nontested subjects, such as social studies and history, we have no way to

validate the instruments outside math and ELA. However, in subjects such as social studies, history, and science, where the nature of effective instruction is plausibly similar to math and ELA, the combination of observations and student survey data could substitute for student achievement gains. We would feel less comfortable extending the analogy to subjects such as art, vocational training, and physical education, where it seems plausible that the model of effective instruction could be very different.

4. **We did not find evidence that any particular set of competencies within the five instruments were more important than the others or that some should be dropped.**

The instruments we tested sought to measure between six and 22 competencies. Partially because of the large-scale scoring effort, FFT, MQI, and PLATO were streamlined to be included in our study. We did not find strong evidence suggesting that these instruments should be pared down further. One challenge is that many of the competencies come together in packages. Much depends on the burden of proof one is using when deciding whether or not to drop a competency. If there is a presumption that any competency that has been nominated by an instructional expert is relevant until proven irrelevant, that will be a high hurdle. We did not find evidence that any of the domains in these instruments are irrelevant. (There are some instruments used by school systems that include many more than 22 competencies. Our results need not imply that these could not be pared down.)

5. **The teachers who perform better on the combination of value-added on state tests, classroom observations, and student feedback *also* have better outcomes on other types of assessments (focused on conceptual understanding and open-ended responses) as well as on student effort and student emotional attachment.**

State tests are limited, imperfect measures. Because of the high cost of scoring open-ended items, state tests tend to focus on multiple-choice items, which may not measure students' conceptual understanding very well. However, it would be a mistake to confuse the nature of the test items with the nature of the instruction associated with gains on those tests. Just because the current state tests are limited does not mean that teachers who succeed on these would not also succeed on richer assessments, if they were available.

By supplementing the state tests with BAM and SAT9 OE, our study provides a rare glimpse into these other outcomes. We find that those teachers who perform better on a combination of classroom observations, student feedback, and value-added (as measured on current state tests) did have students who outperformed on these other assessments. Moreover, their students were more likely to report high levels of effort and a positive emotional association to being in the teacher's class.

6. Whenever multiple measures are being used, policymakers must choose how much weight to attach to each. There are trade-offs, in terms of both validity (in predicting other outcomes) and reliability.

We compared two different approaches to weighting: the equal-weighting approach and the criterion-based approach, which optimizes the weights for predicting student achievement gains (or potentially other outcomes). The criterion-weighted measures placed greater emphasis on value-added than on classroom observations or student feedback. As a result, there were some gains in terms of correlation with underlying value-added.

However, there were also costs: The criterion-weighted measures resulted in smaller differences in student effort and students' positive emotional attachment to class. Moreover, the equally weighted measures seemed to be more reliable. There was little consequence in terms of ability to predict student achievement gains on the supplemental assessments.

7. Current state ELA tests are not as sensitive to teaching effects as the SAT9 OE.

Like many others, we find evidence that the heterogeneity in teacher effects on state ELA tests is considerably smaller than in mathematics. However, our results raise questions about whether this familiar result is simply due to the nature of the state ELA assessments today, which almost universally take the form of multiple-choice reading comprehension tests. Assessments that require students to write may well be more sensitive to the work teachers do in literacy outside the early grades, where reading comprehension is the primary focus.

Three Key Take-Aways

The MET project is in many ways unprecedented: its large scale, its use of multiple indicators and alternative student outcomes, and its random matching of teachers to classrooms in the second year of the study. We emphasize three key points we hope readers will take away from this report.

High-quality classroom observations will require clear standards, certified raters, and multiple observations per teacher.

Clear standards and high-quality training and certification of observers are fundamental to increasing inter-rater reliability. However, when it comes to measuring consistent aspects of a teacher's practice, reliability will require more than inter-rater agreement on a single lesson. Because teaching practice varies from lesson to lesson, multiple observations will be necessary when high-stakes decisions are to be made. But how will school systems know when they have implemented a fair system? Ultimately, the most direct way to do so is to periodically audit a representative sample of official observations, by having impartial observers perform additional observations. In the report, we describe one approach to doing this.

Combining the three approaches (classroom observations, student feedback, and value-added student achievement gains) capitalizes on their strengths and offsets their weaknesses.

For example, value-added is the best single predictor of a teacher's student achievement gains in the future. But value-added is often not as reliable as some other measures and it does not point a teacher to specific areas needing improvement. Classroom observations provide a wealth of information that could be used to support teachers in improving their practice. But, by themselves, they're not highly reliable, and they are only modestly related to student achievement gains. Student feedback promises greater reliability because it includes many more perspectives based on many more hours in the classroom, but it is not as highly related to achievement gains as another value-added measure would be. Each shines in its own way, either in terms of predictive power, reliability, or diagnostic usefulness.

Combining new approaches to measuring effective teaching—while not perfect—significantly outperforms traditional measures. Providing better evidence should lead to better decisions.

No measure is perfect. But if every personnel decision carries consequences—for teachers and students—then school systems should learn which measures are better aligned to the outcomes they value. Combining classroom observations with student feedback and student achievement gains on state tests did a better job than master's degrees and years of experience in predicting which teachers would have large gains with another group of students. But the combined measure also predicted larger differences on a range of other outcomes, including more cognitively challenging assessments and student-reported effort and positive emotional attachment. We should refine these tools and continue to develop better ways to provide feedback to teachers. In the meantime, it makes sense to compare measures based on the criteria of predictive power, reliability, and diagnostic usefulness.

Future Analyses

Our next report, in mid-2012, will evaluate the 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. In this paper, we discussed two approaches: equal weighting and criterion-based weighting. In our next report, we will explore the rationale for and consequences of a variety of different weighting strategies.

The most vexing question we face is whether or not any of the results above were biased by student characteristics not included in our statistical controls. 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, etc.

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 later this year, we will report those findings.

In our final report, we will also incorporate several additional measures. 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.

Referring to the report in mid-2012 as our "final" report 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—such as studying differences in impacts on different subgroups of students, scoring the MET project videos using other (perhaps "second generation") instruments and testing their validity and reliability, and looking for interactions among the different competencies from each of the instruments.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). "Teachers and student achievement in the Chicago public high schools." *Journal of Labor Economics* 25(1): 95–135.

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011) "An interaction-based approach to enhancing secondary school instruction and student achievement." *Science* 333(6045):1034–37.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010) "Problems with the use of student test scores to evaluate teachers." Economic Policy Institute briefing paper no. 278.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Brown Center on Education Policy, The Brookings Institution, November 17.

Goldhaber, D., & Hansen, M. (2010). "Is it just a bad class? Assessing the stability of measured teacher performance." Center for Education Data and Research, working paper #2010-3.

Harris, D., & Sass, T.R. "Value-added models and the measurement of teacher quality" (unpublished paper, April 3, 2006). Available at http://eps.education.wisc.edu/Faculty%20papers/Harris/Manuscripts/IES%20Harris%20Sass%20Value-added%20142.pdf.

Hill, H. (2009). Evaluating value-added models: A measurement perspective. *Journal of Policy Analysis and Management* 28: 702–709.

Kane, T.J. (2004). "The impact of after-school programs: Interpreting the results of four recent evaluations" working paper. New York: William T. Grant Foundation. Available at www.wtgrantfoundation.org/publications_and_reports/browse_reports/kane_working_paper.

Koedel, C., & Betts J.R. (2007). "Re-examining the role of teacher quality in the educational production function." Working paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., & Mihaly, K. (2009). "The intertemporal variability of teacher effect estimates." *Education Finance and Policy* 4: 572–606.

Neal, D.A., & Johnson, W.R. (1996). "The role of premarket factors in black-white wage differences." *Journal of Political Economy* 104(5): 869–95.

Novick, M.R. (1966). "The axioms and principal results of classical test theory." *Journal of Mathematical Psychology* 3(1): 1–18.

Sass, T.R. (2008). "The stability of value-added measures of teacher quality and implications for teacher compensation policy." Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, the Urban Institute. Available at www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf.

Sturman, M.C., Cheramie, R.A., & Cashen, L.H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology* 90: 269–83.

Taylor, E.S., & Tyler, J.H. (2011). "The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers." National Bureau of Economic Research working paper no. 16877.

Appendices

Appendix Table 1. Sample Restrictions Leading to Video Sample

	ADJUSTMENT	SAMPLE SIZE
Total MET Project Teachers		2,937
Exclude Pittsburgh (which was a pilot site, but which did not implement full MET project measures)	-196	2,741
Exclude 9th grade (no prior year value-added measures for most measures; will use in final report)	-715	2,026
Exclude 1st year participants that did not participate in randomizaton	-838	1,188
Exclude 1st year randomized sample without useable video scores	-7	1,181
Add 1st year participants with multiple course sections or complete data with videos scored during "Plan B"	+152	1,333

Note: The above table describes sample exclusions leading to the analytic sample for this report. Pittsburgh piloted several of the measures before the other districts, but we never planned to collect the full set of measures there. We excluded the 9th grade sample from this report because we had no prior measure of value-added. That is, many of the districts did not have end-of-course tests in high school during the 2008-09 school year. In the Spring of 2011, we had begun scoring 1st year videos for teachers with complete data, regardless of their participation in the randomization sample (the "Plan B" sample). However, cost pressures and delays in scoring led us to narrow the scoring effort to the subset of first year participants who agreed to randomization.

Plan B: Teachers were deemed eligible for the Plan B sample if they had "complete data," defined as scores on all relevant assessments, value-added measures in both years (2008–09 and 2009–10), student perception data, and complete video data (four videos in a given subject in middle school grades and eight videos in self-contained classrooms). Teachers were selected at random from the eligible pool, after stratifying by district, grade level, and subject.

Appendix Table 2. Regression Coefficients Used for Criterion-Based Weighting

	OUTCOME FOR REGRESSION:	
MEASURES FROM 2010 USED TO PREDICT GAINS IN ANOTHER CLASSROOM	MATH GAIN IN OTHER CLASSROOM	ELA GAIN IN OTHER CLASSROOM
FFT		
FFT	0.042	0.039
Tripod	0.200	0.294
State test	0.758	0.666
CLASS		
CLASS	0.092	-0.038
Tripod	0.179	0.330
State test	0.729	0.708
UTOP		
UTOP	0.047	
Tripod	0.313	
State test	0.640	
MQI		
MQI	0.039	
Tripod	0.294	
State test	0.666	
PLATO		
PLATO		0.176
Tripod		0.246
State test		0.578

Note: The above are regression coefficients, using student achievement gains in another course section as the dependent variable, and teachers' scores on the observation instrument, Tripod and gains on the state test from another class as regressors. These are the weights used to construct the combined measures in Tables 16 and 20.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org