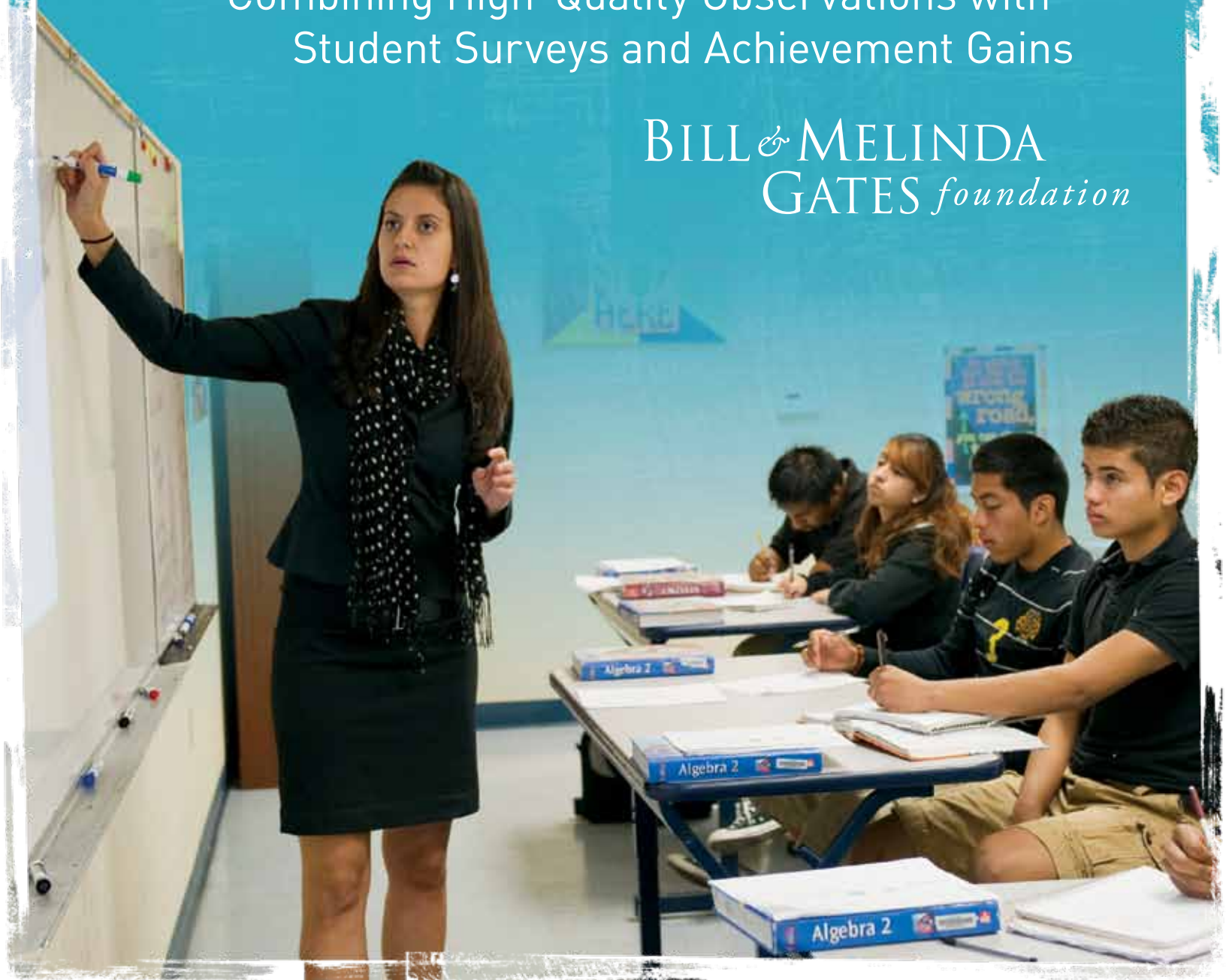
MET project

POLICY AND PRACTICE BRIEF

Gathering Feedback for Teaching

Combining High-Quality Observations with Student Surveys and Achievement Gains

BILL & MELINDA GATES foundation



ABOUT THIS REPORT: This report is intended for policymakers and practitioners wanting to understand the implications of the Measures of Effective Teaching (MET) project's interim analysis of classroom observations. Those wanting to explore all the technical aspects of the study and analysis also should read the companion research report, available at www.metproject.org.

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 using random assignment to study the extent to which student assignment may affect teacher effectiveness results.

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
- 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

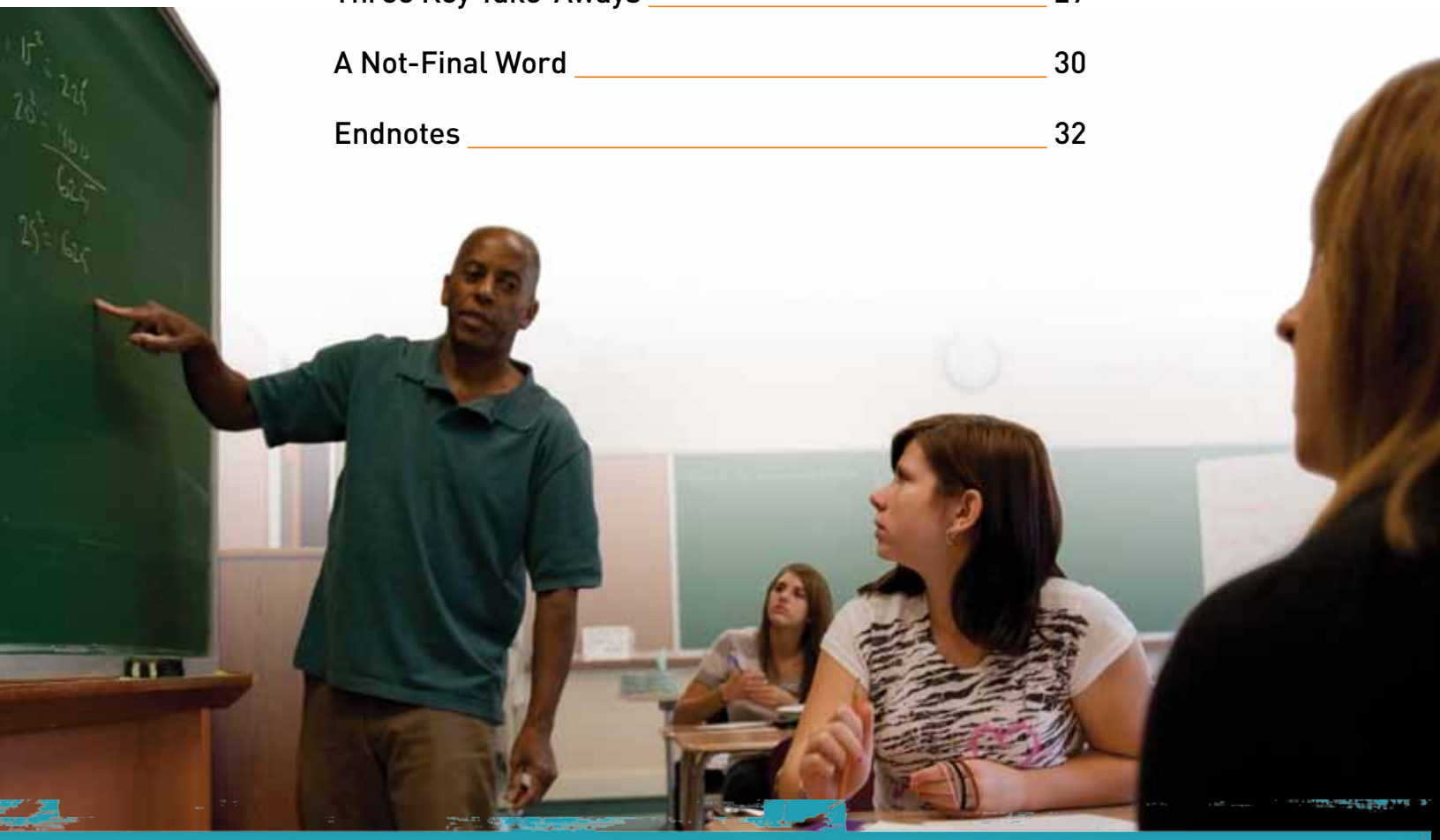
Acknowledgments: In addition to MET project partners who reviewed early drafts of the companion research report, we would like to thank the following external experts who read and provided written feedback on the document: Anthony Bryk, Andrew Ho, Bob Linn, Susan Moore Johnson, and Jonah Rockoff. The lead authors accept full responsibility for any remaining errors in the analysis.[1]

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.

The 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.

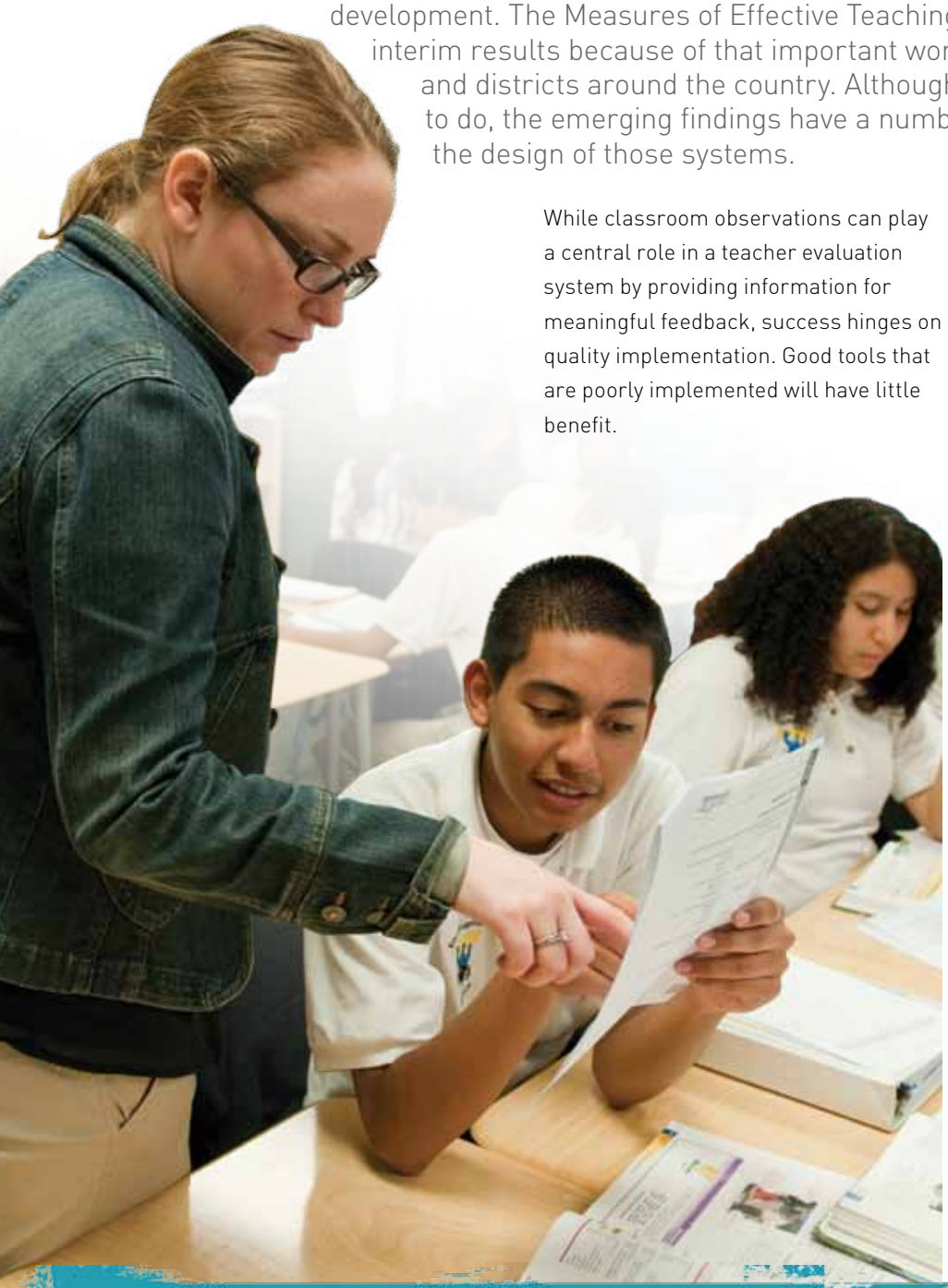
January 2012

Contents



Guidance to Policymakers and Practitioners

Policymakers and practitioners at every level are intensely focused on improving teaching and learning through better evaluation, feedback, and professional development. The Measures of Effective Teaching (MET) project is releasing these interim results because of that important work already under way in states and districts around the country. Although the project has much work still to do, the emerging findings have a number of important implications for the design of those systems.



While classroom observations can play a central role in a teacher evaluation system by providing information for meaningful feedback, success hinges on quality implementation. Good tools that are poorly implemented will have little benefit.

Therefore, we emphasize the following six minimum requirements for high-quality classroom observations:[2]

1. **Choose an observation instrument that sets clear expectations.** That means defining a set of teaching competencies and providing specific examples of the different performance levels on each. Many such instruments are already available and will be improving over time. Lengthy lists of vaguely described competencies are not sufficient.

2. **Require observers to demonstrate accuracy before they rate teacher practice.** Teachers need to know that observers can apply an observation instrument accurately and fairly—*before* performing their first observation. Good training is not enough. Observers should be expected to demonstrate their ability to generate accurate observations and should be recertified periodically.

3. **When high-stakes decisions are being made, multiple observations are necessary.** For teachers facing high-stakes decisions, the standard of reliability should be high. Our findings suggest that a single observation cannot meet that standard. Averaging scores over multiple lessons can reduce the influence of an atypical lesson.

4. **Track system-level reliability by double scoring some teachers with impartial observers.** At least a representative subset of teachers should be observed by impartial observers with no personal relationship to the teachers. This is the only way to monitor overall system reliability and know whether efforts to ensure reliability are paying off.

5. **Combine observations with student achievement gains and student feedback.** The combination of classroom observations, student feedback, and student achievement carries three advantages over any measure by itself: (a) it increases the ability to predict if a teacher will have positive student outcomes in the future, (b) it improves reliability, and (c) it provides diagnostic feedback that a teacher can use to improve. In the grades and subjects where student achievement gains are not measured, classroom observations should be combined with student feedback surveys.

6. **Regularly verify that teachers with stronger observation scores also have stronger student achievement gains on average.** Even a great observation instrument can be implemented poorly. And any measure can become distorted in use. (This could be true for student feedback surveys as well.) Rather than rely on this study or any other as a guarantee of validity, school systems should use their own data to confirm that teachers with higher evaluation scores also have larger student achievement gains, at least *on average.*

Executive Summary

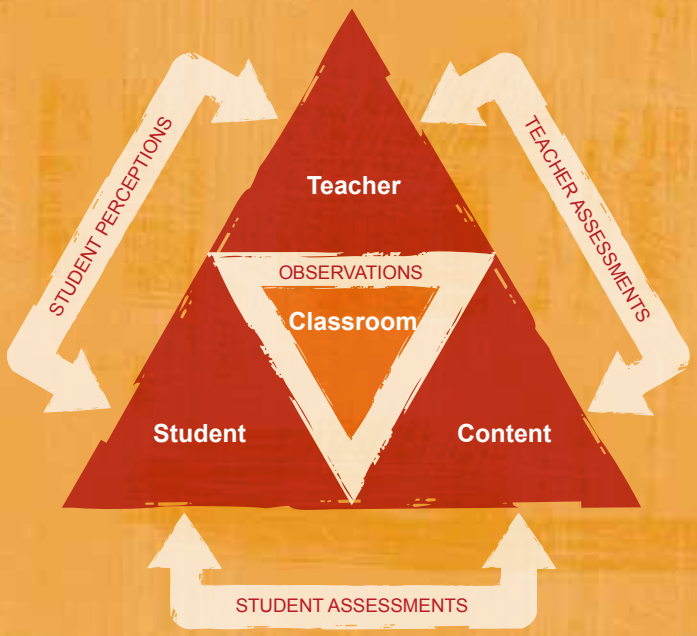
Research has long been clear that teachers matter more to student learning than any other in-school factor. Improving the quality of teaching is critical to student success. Yet only recently have many states and districts begun to take seriously the importance of evaluating teacher performance and providing teachers with the feedback they need to improve their practice.

The MET project is working with nearly 3,000 teacher-volunteers in public schools across the country to improve teacher evaluation and feedback. MET project researchers are investigating a number of alternative approaches to identifying effective teaching: systematic classroom observations; surveys collecting confidential student feedback; a new assessment of teachers' pedagogical content knowledge; and different measures of student achievement. See **Figure 1**.

Figure 1

MET Project

Multiple Measures of Teaching



In a previous paper, we reported that confidential student surveys about students' classroom experiences can provide reliable and meaningful feedback on teaching practice.[3] In this report, we investigate the properties of the following five instruments for classroom observation:

- **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).

All the instruments establish a set of discrete competencies and then describe observable indicators of different levels of performance. We studied each instrument using two criteria:

1. **Reliability.** Reliability is the extent to which results reflect consistent aspects of a teacher's practice and not the idiosyncrasies of a particular observer, group of students, or lesson.

2. **Validity.** Validity is the extent to which observation results are related to student outcomes.

If any of the instruments listed is to be helpful in practice, it will need to be implementable at scale. To that end, our analysis is based on 7,491 videos of instruction by 1,333 teachers in grades 4–8 from the following districts: Charlotte-Mecklenburg, N.C.; Dallas; Denver; Hillsborough Co., Fla.; New York City; and Memphis.[4] Teachers provided video for four to eight lessons during the 2009–10 school year. Some 900 trained raters took part in the subsequent lesson scoring. We believe this to be the largest study ever to investigate multiple observation instruments alongside other measures of teaching.

Key Findings:

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

   The teachers who more effectively demonstrated the types of practices emphasized in the instruments had greater student achievement gains than other teachers.

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

   In our study, the same teacher was often rated differently depending on who did the observation and which lesson was being observed. The influence of an atypical lesson and unusual observer judgment are reduced with multiple lessons and observers.

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

   Observations alone, even when scores from multiple observations were averaged together, were not as reliable or predictive of a teacher's student achievement gains with another group of students as a measure that combined observations with student feedback and achievement gains on state tests.

4. **Combining observation scores, student feedback, and student achievement gains was better than graduate degrees or years of teaching experience at predicting a teacher's student achievement gains with another group of students on the state tests.**

   Whether or not teachers had a master's degree or many years of experience was not nearly as powerful a predictor of a teacher's student achievement gains on state tests as was a combination of multiple observations, student feedback, and evidence of achievement gains with a different group of students.

5. **Combining observation scores, student feedback, and student achievement gains on state tests also was better than graduate degrees or years of teaching experience in identifying teachers whose students performed well on other measures.**

   Compared with master's degrees and years of experience, the

combined measure was better able to indicate which teachers had students with larger gains on a test of conceptual understanding in mathematics and a literacy test requiring short written responses.

In addition, the combined measure outperformed master's and years of teaching experience in indicating which teachers had students who reported higher levels of effort and greater enjoyment in class.

The following pages discuss the instruments, scoring process, findings, and implications in greater detail. Sections are organized around the elements of the "Pathway to High-Quality Classroom Observations" in **Figure 2** below.

Figure 2

Pathway to High-Quality Classroom Observations
as Part of a Multiple Measures System

Defining Expectations for Teachers

Teachers, supervisors, and providers of professional development need a common vision of effective instruction to work toward. For this reason, the observation instruments in the MET project are not checklists focusing on easy-to-measure but trivial aspects of practice, such as whether or not lesson objectives are posted. Rather, for each defined competency the instruments require judgment about how well the observed practice aligns with different levels of performance. This is illustrated by the excerpts below from FFT.

Indicators of Cognitive Challenge in "Use of Questioning and Discussion Techniques" from FFT

UNSATISFACTORY	BASIC	PROFICIENT	DISTINGUISHED
Low cognitive challenge, predominantly recitation.[5] *Ex: Teacher points to PowerPoint slide and asks: "What does this say?"*	**Some** questions reflect **moderate cognitive challenge.** *Ex: Teacher asks mix of higher-order questions and questions with single correct answers.*	**Variety of questions** challenge students and **advance high-level thinking/discourse.** *Ex: Most of teacher's questions are open-ended, as, "What might have happened if the colonies had not prevailed in the American War for Independence?"*	In addition to indicators in proficient column, **students initiate higher-order questions.** *Ex: A student asks of other students, "Does anyone have another idea as to how we might figure this out?"*

Each instrument embodies a particular vision of effective instruction, reflected in the set of competencies on which it chooses to focus attention and how it defines proficiency in those competencies.[6] The challenge for any instrument developer is to identify a manageable set of competencies and describe them with sufficient specificity to allow observers to score reliably. To address this, the five instruments in the MET project take varied approaches to the total number of competencies and performance levels, as outlined in the table on the following page.

Some of the instruments were streamlined for the study, given the MET project's decision to score on a large scale using video. For example, the version of FFT used in the study lacks two competencies—or "components" as FFT calls them—found in the full version of the instrument: Flexibility and Responsiveness & Use of Physical Space. We determined that these would be difficult to score accurately without conferring with the teacher about issues such as how the lesson was planned based on student understanding and whether the teacher was teaching in his or her regular classroom. Similarly, PLATO includes as many as 13 competencies, but the PLATO Prime instrument that the MET project scored includes six. The version of MQI used, MQI Lite, also included six competencies, while the full version subdivides those six into 24 elements that each receives its own score.

Diagnosing Teacher Practice

The different instruments paint a similar portrait of the nature of teaching in the MET project classrooms. The potential that observations hold for providing diagnostic information is evident in the distributions of scores of the MET project volunteers. Observers used multiple measures to score the same 30 minutes of instruction, and observers scored them in segments that varied according to each instrument's guidelines (15 minutes for CLASS and 7.5 minutes for MQI, for example). **Figure 3,** on pages 10–11, shows the distribution of scores given to lessons taught by MET project volunteers for each competency on each of the given instruments.

Two patterns are clear in the study sample. First, overall observed practice is overwhelmingly in the mid-range of performance as defined by the instruments. Second, scores are highest for competencies related to creating an orderly environment and lowest for those associated with the most complex aspects of instruction. On FFT, for example, more than two-thirds of scores given for "managing student behavior," "creating an environment of respect and rapport," and "engaging students in learning" are proficient or above. But the proficiency-and-above rate is just 44 percent for scores on "using assessment in instruction," 34 percent for "using questioning and discussion techniques," and 30 percent for "communicating with students" (the last competency requires clear presentation of content as well as culturally and developmentally appropriate communication). Yet it's these kinds of more complex teaching skills that will be required for success on the new Common Core State Standards.

The classroom observation's greatest promise lies in its use as a developmental tool. To realize that promise, professional development will need to be individualized to meet teachers' specific needs (just as content is being individualized to meet students' needs in some schools today). The MET project has not yet investigated the impact of such new models of professional development, which are explicitly aligned with teachers' evaluation results. However, there is encouraging evidence emerging from other research suggesting that such individualized feedback to teachers can lead to better outcomes for students.[7]

The MET Project's Observation Instruments[8]

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

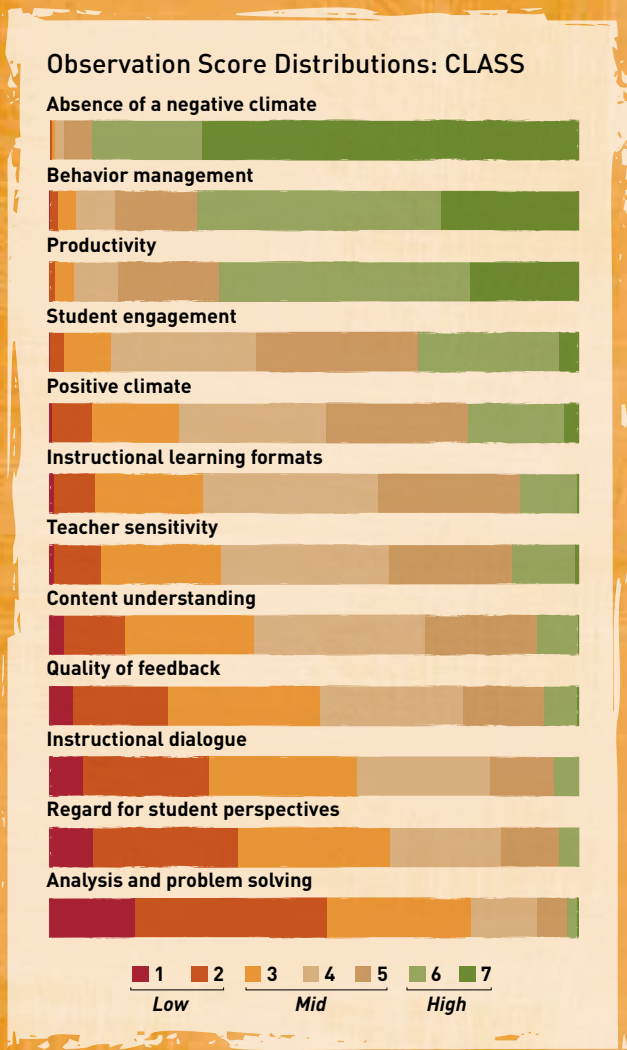
Figure 3

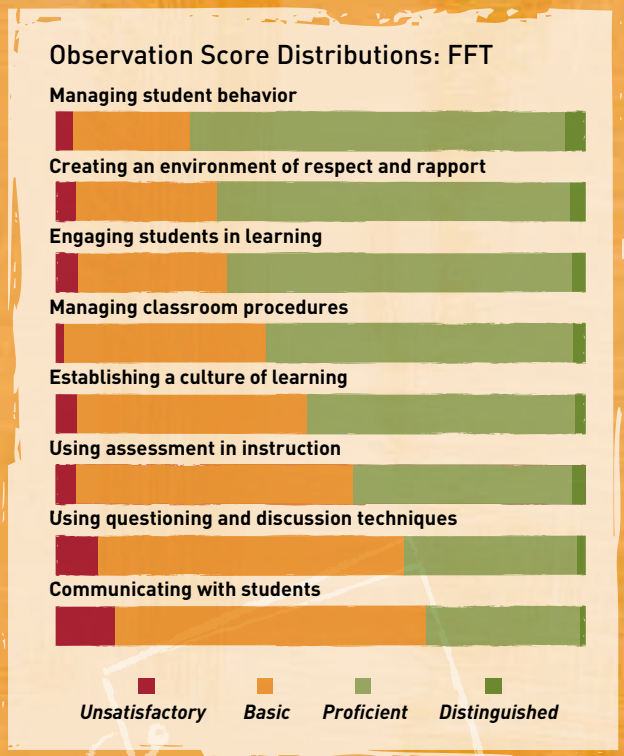
Observing Teaching Practice

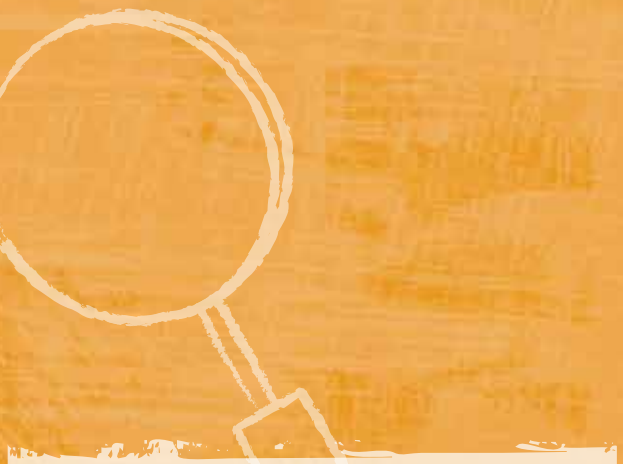
Through Five Lenses

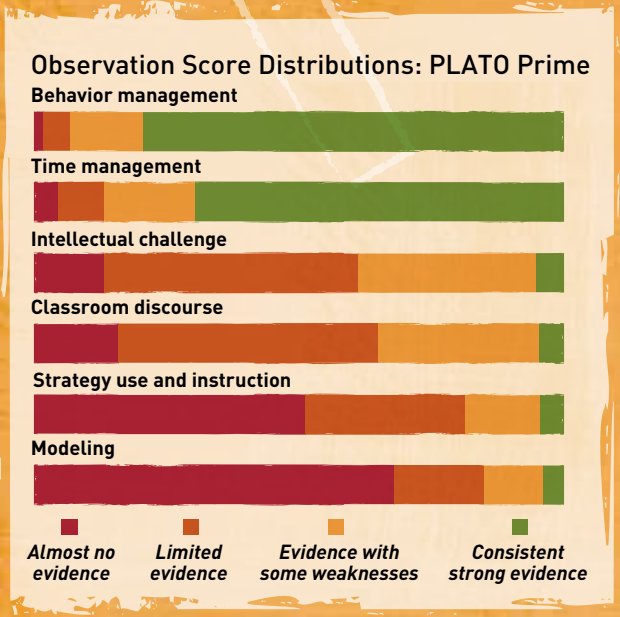
Each chart represents one of the five observation instruments used in the MET project to assess classroom practice. Each row represents the performance distribution for a particular competency. All of the instruments use scales with between three and seven performance levels. Higher numbers represent 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.

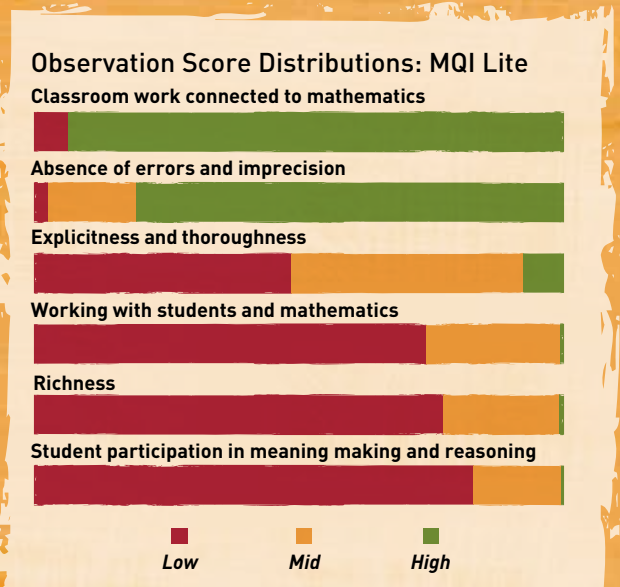
A few patterns are immediately visible in the data. First, raters judged the observed lessons to be orderly and generally on-topic. Across these instruments, behavioral-, time-, and materials-management competencies were rated as most accomplished. Second, across all instruments, raters rarely found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

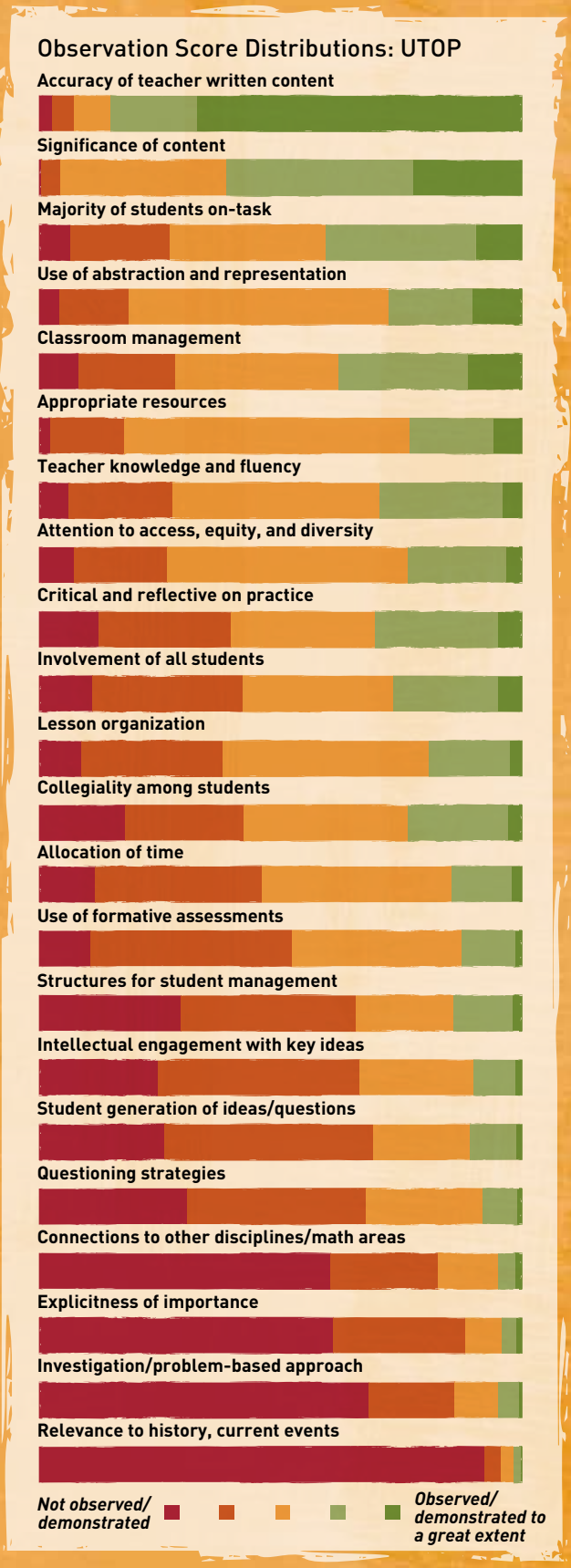







Observation Score Distributions: PLATO Prime
Behavior management
Time management
Intellectual challenge
Classroom discourse
Strategy use and instruction
Modeling
Almost no evidence
Limited evidence
Evidence with some weaknesses
Consistent strong evidence


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

Which Competencies Most Relate to Student Achievement?

While some of the instruments studied in the MET project were streamlined for the study, it is possible that they could be simplified further and still provide reliable and useful feedback. One approach would be to prioritize competencies based on their relationship with student achievement outcomes and eliminate those most weakly related. Indeed, practitioners frequently ask the MET project which competencies most predict achievement gains because they are looking for rationales to guide instrument refinement.

Answering the question is harder than it seems. It starts by finding which competencies, or groups of competencies, are independent from each other and which tend to cluster together. If teachers who rank a certain way on one competency always rank a certain way on another, then it would be impossible to determine which of the two competencies is most associated with outcomes because they always occur (or fail to occur) together. To know how important a given competency is, there must be some degree of independence between that competency and the others. Only if there are sufficient numbers of teachers with unexpected combinations of skills will a school system be able to measure whether any given competency matters, while holding all others constant.

"Clusterings" of Competencies

In examining the scores of MET project volunteers on all of the instruments, we found three "clusterings" of competencies. For all five instruments, the first cluster included all competencies. That is, those who scored well on one competency tended to score well on all others. For all five instruments, a second cluster included competencies related to classroom and time management. On FFT, for example, teachers who ranked highly on "managing classroom procedures" almost always ranked highly on "managing student behavior," and vice versa. A third clustering reflected something unique about the instrument's core instructional approach. With CLASS, for example, this cluster included competencies associated with teacher sensitivity and the affective climate in the classroom. In FFT, this cluster focused on a teacher's ability to elicit student thinking through questioning and assessment skills. Indeed, we could not conclude that any of the clusters were unrelated to student achievement gains. In fact, in most cases we could not conclude that any cluster was more important than the others. Further research is needed to determine why competencies cluster: Do they actually occur in predictable patterns in practice, or is it too hard for observers to keep them separate? We will continue to study strategies for instrument refinement.

Relationships across Instruments

We did find notable patterns in the relationships among the different instruments. Teachers' scores on FFT and CLASS were highly correlated, suggesting that the two cross-subject instruments measure very similar things, or at least measure competencies that almost always occur together. Likewise, scores on the two math instruments—UTOP and MQI—were highly related. However, the correlations between the cross-subject and math instruments were lower. The math instruments seem to be measuring a somewhat different set of skills. This was not the case with PLATO, the ELA-specific instrument, which showed much more association with the cross-subject instruments than did the math instruments.

Challenges & Considerations for Defining Expectations for Teachers

Challenge

School systems must adopt, adapt, or create observation instruments that match their theory of instruction. In doing so they face a balancing act: The more comprehensive and detailed the instrument, the more likely it is to encapsulate a vision of effective instruction, but at some point the tool can become so complex that it overloads observers and makes it impossible for individuals to render judgments of practice according to guidelines.

Considerations

Are there redundancies that can be eliminated? Looking at the indicators underneath each competency may reveal overlaps that suggest the total number of competencies can be collapsed without sacrificing the overall instructional construct.

How many performance levels are needed? To provide information for improving practice, instruments must describe practice along a continuum of performance. But if observers find it too difficult to discern between adjacent levels of performance, school systems should consider reducing the number of performance categories.

Can language be clarified? If well-trained observers find it difficult to use an instrument, one reason may be that its competencies are not defined clearly enough. That said, clarifying competencies is another balancing act; at some point definitions can become so objective they result in too rigid an interpretation of good instruction.

Examples in Practice

Washington, D.C. In the District of Columbia Public Schools, district leaders revised the system's observation instrument, called the Teaching and Learning Framework, following the first year of implementation. Many of the changes were aimed at supporting observers in making more accurate judgments. Among them:

1. Clarifying language, as in revising a standard from "Engage all students in learning" to "Engage students at all learning levels in rigorous work."

2. Allowing greater flexibility where needed, such as revising an indicator that specified teachers should target three learning styles within 30 minutes of instruction to instead emphasize providing students with multiple ways to engage with the content.

3. Collapsing the total number of standards by identifying overlaps so that different standards related to classroom management and productivity were brought under a single expectation, "Maximize Instructional Time."

Ensuring Accuracy of Observers

Inaccurate classroom observations lead to mistrust and poor decisions. Ensuring accuracy is not just a matter of training. It requires assessing observers' ability to use the instrument at the end of training. Moreover, it may be necessary to ask observers to redemonstrate periodically their ability to score accurately. In the MET project, our partners the Educational Testing Service (ETS) and Teachscape jointly managed the recruitment and training of observers (or "raters" as we called them) and lesson scoring for four of the five instruments. For UTOP, the National Math and Science Initiative managed scoring.

Most training and all scoring was conducted online. All raters held a bachelor's degree and a majority (about 70 percent across most instruments) held higher degrees. Some were currently enrolled in teacher preparation programs, but the vast majority (more than 75 percent) had six or more years of teaching experience.

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.

At the end of their training, raters were required to rate a number of videos pre-scored by experts and achieve a minimum level of agreement with the expert scores.[9] Raters who failed to meet this certification standard 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 4**). The pass rate for raters averaged 77 percent across instruments.

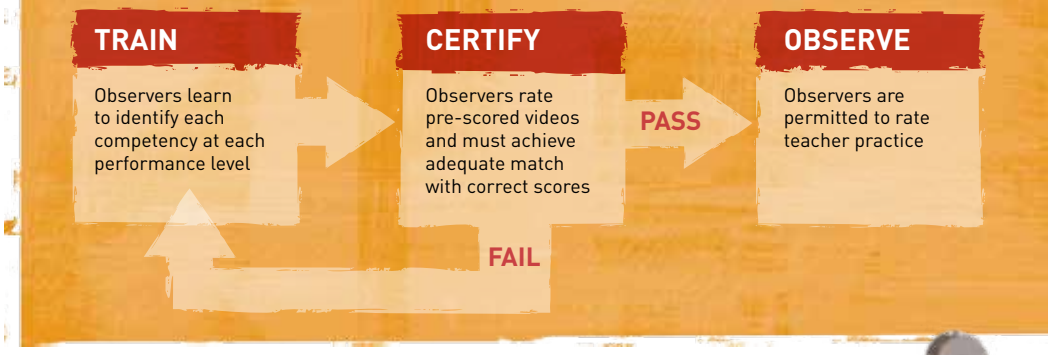
The MET project also monitored rater accuracy on an ongoing basis. At the start of each shift, raters had to pass a "calibration" assessment, scoring a smaller set of pre-scored videos. Raters were given two opportunities to meet

the calibration standard at the start of each shift. Those who did not meet the standard after both tries (about 10 percent on average) were not permitted to score videos that day. Raters who failed calibration received additional training and guidance from their "scoring leader"—an expert scorer responsible for managing and supervising a group of raters.

In addition, pre-scored videos were interspersed with the unscored videos assigned to each rater (although raters were not told which were pre-scored). 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 submitted discrepant scores on the pre-scored videos.

Figure 4

Ensuring Accuracy of Observers



Another quality control check was double scoring, in which the same video was scored by two different raters so the results could be compared.

A MET project teacher captures a lesson using the Teachscape panoramic camera.

Challenges & Considerations for Ensuring Accuracy of Observers

Challenge

Observer training doesn't ensure quality. Only by having observers demonstrate their ability to score with adequate accuracy can school systems know if their training has been successful or if it needs to be improved. Assessing the accuracy of observers requires setting standards and creating a process with which to test observers against them.

Considerations

How should "correct" scores be determined? Assessing observer accuracy through scoring videos requires knowing what the right scores are. Outside consultants with expertise in widely available instruments may be able to provide pre-scored, "master-coded" videos. Systems that create customized observation instruments must decide who can best determine correct scores and how. Involving classroom teachers in master coding may build credibility. Supervisors add the perspective of those who will ultimately make personnel decisions.

The MET project is working with school districts to develop a certification tool that will allow systems to master code MET project videos of teachers (who have given special consent) according to their own instruments and then use the coded videos to assess whether trainees can score accurately before completing their training.

What does it mean to be accurate? Systems must define accuracy before they can assess it. In the MET project, performance standards for raters were determined by the instrument developers and ETS. "Accurate" typically meant achieving a minimum percentage of scores that exactly matched the correct scores (50 percent for FFT) or that were not more than one point off (70 percent for CLASS). Some also specified a maximum percentage of scores that could be "discrepant"—that is, two or more points off from the correct score (10 percent for PLATO).

How often should observers be retested? Observer accuracy may slide after initial training and certification. In determining how often to reassess their observers, districts will need to weigh the costs of calibration against the benefits in terms of accuracy and trust.

Examples in Practice

Memphis. In the Memphis City Schools, training on the district's Teacher Effectiveness Measure (TEM) observation rubric starts with two days on the instrument and how to use it, after which observers practice scoring independently over approximately three weeks.

Observers then take a three-hour refresher session and a certification assessment in which they rate a set of pre-scored video vignettes using the TEM instrument. To pass certification, observers must have scores that exactly match the correct scores for three of seven competencies in the rubric. In addition, for no competency may they deviate by more than one point from the correct score on the tool's five-point scale.

Correct scores for videos are determined by a 17-member certification committee of teachers, principals, and district-level administrators who review videos and assign ratings by consensus before they are used for training and assessment.

Ensuring Reliable Results

Observation results should capture consistent qualities of a teacher's practice. This is why districts often focus on "inter-rater" reliability; they want a teacher's rating to be due to the quality of the lesson and not the quality of the observer. But inter-rater reliability focuses on just one of the reasons—the rater—for why a single observation could be a misleading indicator of a teacher's actual practice. Many other factors may influence ratings, such as the content of a given lesson or the makeup of a particular group of students.

To investigate reliability in our results, we analyzed data from a subset of lessons scored by more than one rater. Because we also had multiple lessons from multiple sections for the same teachers, this let us study the degree to which observation scores varied from teacher to teacher, section to section, lesson to lesson, and rater to rater. We did so in a way that allowed a comparison of how different *sources* of variability in scores affected overall reliability. This is how those sources compared:

- **Teacher effects:** A reliable measure is one that reveals consistent aspects of a teacher's practice. Yet only 14 percent to 37 percent of the variation in overall scores across all the lessons was due to consistent differences among teachers. In other words, a single observation score is largely driven by factors other than consistent aspects of a teacher's practice.
- **Lessons:** For most of the instruments, the variance in scores between lessons for a given teacher was at least half as large as the teacher effect. In other words, even if we had a very precise measure of the quality of instruction in *one lesson*, we would still have had an inaccurate impression of a teacher's practice—because a teacher's score varied considerably from lesson to lesson.

- **Course section:** The particular students sitting in the classroom played little role in the scores. In all five instruments, 4 percent or less of the variation in the overall score was associated with course section. A teacher's score may have varied from lesson to lesson, but it made little difference if those lessons were with different course sections (which also meant the time of day mattered little).

- **Raters:** For most of the instruments, no more than 10 percent of the total variance in scores was due to some raters consistently scoring high and other raters consistently scoring low. In other words, there was little evidence that a large number of raters were "too easy" or "too difficult." Statisticians refer to this as the "main rater effect." (This just means that the average score raters gave across a large number of lessons was not dramatically different. This does **not** mean that on any particular lesson two different raters were in agreement. They often were not.)

- **Unexplained effects:** For every instrument, the largest source of variance fell into this final category, what researchers call "the residual variance." For example, when different raters watched a given lesson and gave it different scores, this would be categorized as residual variance. While few raters showed a consistent tendency to be "too hard" or "too lenient" across a large number of lessons, they often disagreed when watching any given lesson.

Achieving High Reliabilities for Our Study

Having determined the extent to which different sources of variance affected scores, we then estimated reliability under different scenarios. In our analysis, the two main obstacles to reliability were:

1. Variation from lesson to lesson, and
2. Differing judgments of raters watching the same lesson.

Averaging over multiple lessons and involving more than one observer can reduce the influence of an atypical lesson and unusual observer judgment. On all the instruments, a single observation by a single rater proved to be a poor indicator of a teacher's typical practice. With two lessons, our reliabilities increased substantially. However, only when we averaged four observations per teacher, each by a different rater, did the variation due to teacher effects account for about two-thirds of the overall variation in observation scores (see **Figure 5**).

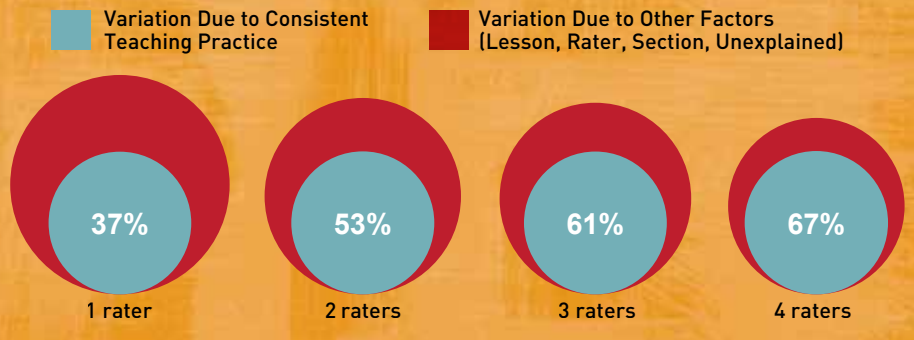
Implications for Practitioners on Ensuring Reliability of Classroom Observations

We caution against concluding from our results that every state or district should require four observations, each by a different rater, to obtain similar reliability. Our study was unique in many ways:

- Our observers were trained and required to demonstrate their ability to score accurately before they could begin scoring;
- Our observers had no personal relationship with the teachers being observed;
- Our observers were watching digital video rather than being present in person; and
- There were no stakes attached to the scores for teachers in our study (or for the observers).

Figure 5

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. These reliability estimates are based on having trained raters, with no personal relationship to teachers, observe digital video. The reliability achieved by school systems could be higher or lower. See Table 11 in the research paper at www.metproject.org for results with the other instruments.

Districts could discover that they require a different number of observations to achieve similar levels of reliability; some may require more than four, others less. Nevertheless, our experience leads us to offer the following guidance to the many states and districts that are redesigning their classroom observations:

- First, we re-emphasize that to achieve high levels of reliability with classroom observations, observers should demonstrate their ability to use the instruments reliably *before* doing an observation. One way to do this would be to use a process similar to that used in the MET project, in which expert observers pre-score a set of videos using a particular observation instrument and then require prospective observers to reproduce those scores by watching those videos online. Another would be to have a trained observer physically accompany each prospective observer on some initial observations and compare notes afterward.

- Second, to produce reliable results on a teacher's practice, districts will need to observe a teacher multiple times. In our study, individual teachers' scores varied considerably from lesson to lesson. This could be for a number of reasons. Different material may require teachers to showcase different skills; no one lesson provides a complete picture of their practice. Or teachers may simply have an off day. Whatever the reason, the same teacher may look different on a different day.

- Third, to monitor the reliability of their classroom observation systems and ensure a fair process, districts will need to conduct some observations by impartial observers. Comparing those scores with scores done by personnel inside the school is the only way to learn whether preconceived notions or personal biases (positive or negative) are driving the scores.[10] Only by double scoring a subset of MET project videos were we able to determine reliability. Likewise, school systems will not know how reliable their observations are without some double scoring.

  The scale of the required double scoring depends on the goal. An initial goal could be to monitor the reliability of the school system as a whole (and not monitor reliability in every single school). This could be done for a representative subset of teachers by drawing a random sample of, say, 100 teachers and having impartial observers conduct an additional observation. (See our research report for more on how such an audit might function.)

After capturing a lesson for the study, a MET project teacher reviews her video. On the left of the screen is the footage from the panoramic camera; on the right is the image from the board camera.

Challenges & Considerations for Ensuring Reliability of Classroom Observers

Challenge

For classroom observations to indicate reliably a teacher's practice requires multiple observations. How many will depend on the quality of observer training and the quality of procedures that observers use for collecting evidence. Regardless, many school systems will need to increase the overall number of observations they perform to achieve reliable results. Doing so with limited resources requires thinking differently about how to allocate time and personnel.

Considerations

Who can share the responsibility of observing? Increasing the number of people who are trained and qualified to observe is one way to boost capacity. Master teachers and instructional coaches from either inside or outside the building could assist. Use of digital video as a complement to in-person observations also would allow multiple observers to view instruction outside the normal school day.

When is reliability most important? A single observation may be adequate to provide informal coaching. But a more reliable and complete picture of a teacher's practice is warranted when observations are part of a formal evaluation that informs high-stakes decisions, such as determining tenure or intervening with a struggling teacher.

How can time be used differently? It may not be necessary for every observation to be equally long or comprehensive. Teachers who have already demonstrated basic skills could be the focus of more targeted observations aimed at higher levels of performance. In addition, systems may get a more complete picture of teacher practice if they have more frequent, shorter observations (ideally, by more than one person), rather than fewer longer ones.

Examples in Practice

Allocating Resources Differently

Denver. The Denver Public Schools varies the number of competencies it includes in each observation. Teachers are observed four times each year, but only in two of those are they observed on all competencies in the system's instrument. During each of the other two, they are observed on two competencies representing specific areas of focus for the teachers.

Hillsborough Co., Fla. Hillsborough has dramatically increased its total number of teacher observations (from as few as one every three years, not long ago), while differentiating its approach to individual teachers. The number of observations teachers receive each year is determined by their prior year's evaluation score. Those who receive the lowest ratings have 11 observations (including formal and informal, administrative, and peer observations). Those with the highest ratings have five.

Like Denver, Hillsborough also varies the length of observations. For top teachers, two of the five observations are of full lessons; the other three are informal observations lasting 20–25 minutes each. For struggling teachers, seven of the 11 observations are formal full-lesson observations, and four are informal.

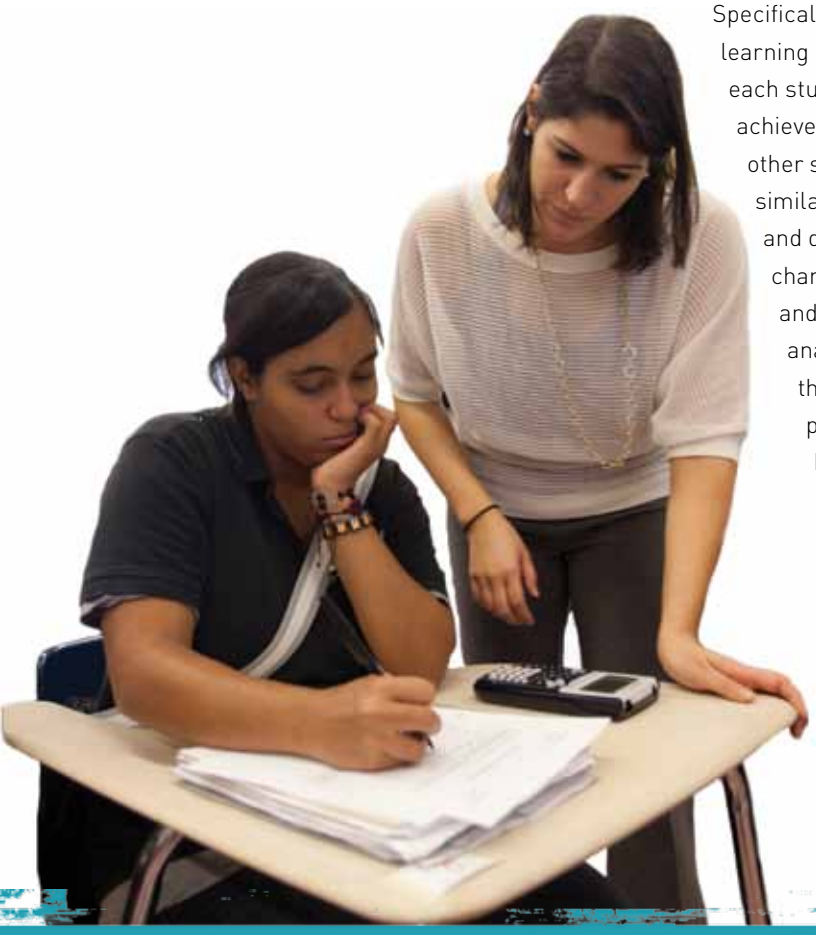
Checking Reliability

Washington, D.C. Teachers in the District of Columbia Public Schools are observed by their principals and by "Master Educators" (MEs). The district has hired and trained about 45 MEs, who go from school to school observing teachers and debriefing with them. Because the MEs are generally drawn from outside a specific school, they can bring a fresh perspective, unburdened by preconceived notions of a teacher's practice (positive or negative). The division of labor enables a comparison of principal and ME scores for each teacher, which the district does to check reliability.

Hillsborough Co., Fla. As noted above, in Hillsborough teachers are observed not just by their principals but also by others: peer evaluators and mentors, both districtwide positions. The school system checks the alignment in final evaluation scores given to individual teachers by principals and the outside evaluators.

Determining Alignment with Outcomes

The ultimate goal is to use classroom observations to help teachers improve their practice and thus student achievement outcomes. Observation instruments that bear no relationship to student outcomes will be of little help in achieving this. To assess the validity of the five instruments, we tested the alignment of each with student achievement gains. We looked at gains rather than end-of-year scores because end-of-year test scores partially reflect differing starting points, whereas we wanted to know if the *progress* students made was related to the teacher's instructional practice. For this reason, we used a "value-added" model to compare teachers in terms of student growth over time.



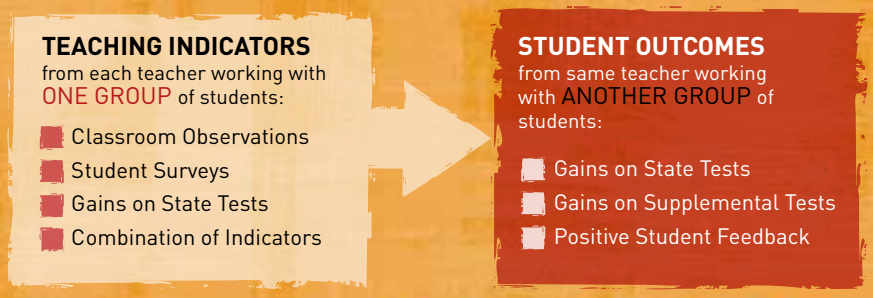
Specifically, we measured learning gains by comparing each student's end-of-year achievement with that of other students who had similar prior performance and demographic characteristics and—because our analysis has shown the importance of peer effects—who had classmates with similar prior performance and demographics. Along with demographics, we controlled for whether students were eligible for free and reduced-price lunches, considered English language learners, or in special education. Based on these comparisons, our statistical model produced a "predicted" achievement score for each student. A teacher's value-added score reflected the difference between his or her students' average predicted scores and their actual ones.

In addition to state tests, students in participating classes took supplemental performance assessments: the Balanced Assessment in Mathematics (BAM) and the Stanford 9 Open-Ended (SAT9 OE) reading assessment.[11] We chose these tests because they included

Figure 6

Testing Validity

To What Extent Do Indicators Predict Outcomes?



cognitively demanding content, were reasonably well aligned with the curriculum in the six states (while including different types of questions than the state tests), had high levels of reliability, and had evidence of fairness to members of different groups of students. Because students only took the supplemental tests once, we generated value-added scores for teachers on the BAM and SAT9 OE using state test scores as the indicator of prior performance.

Testing Alignment across Different Groups of Students

Testing for validity means determining the extent to which teaching indicators are related to student outcomes. In the MET project, we tested this relationship by comparing a teacher's results on teaching indicators from working with *one* group of students to outcomes from the same teacher working with *another* group of students. For example, did a teacher who had high value-added scores and positive student feedback in one class produce high value-added scores with a different class?

We did this for two reasons. In some cases we wanted to know the extent to which a teacher's results on the same measure were similar when working with different groups of students. The second reason related to the testing of classroom observations. Observer judgments may have been biased by the student behaviors they saw, which could also be related to student achievement gains.[12] To address both concerns, we compared a teacher's observation scores from working with one group of students to the value-added scores of the same teacher working with a different group of students (see **Figure 6**).

We did this in two ways. For teachers who taught two sections of the same class during the year, we collected data, and we compared the indicators and outcomes from those two sections. For those who taught self-contained classes, we compared their observation scores (and other indicators, such as student surveys) from the year they were videoed to their value-added scores for the same grade the previous year.

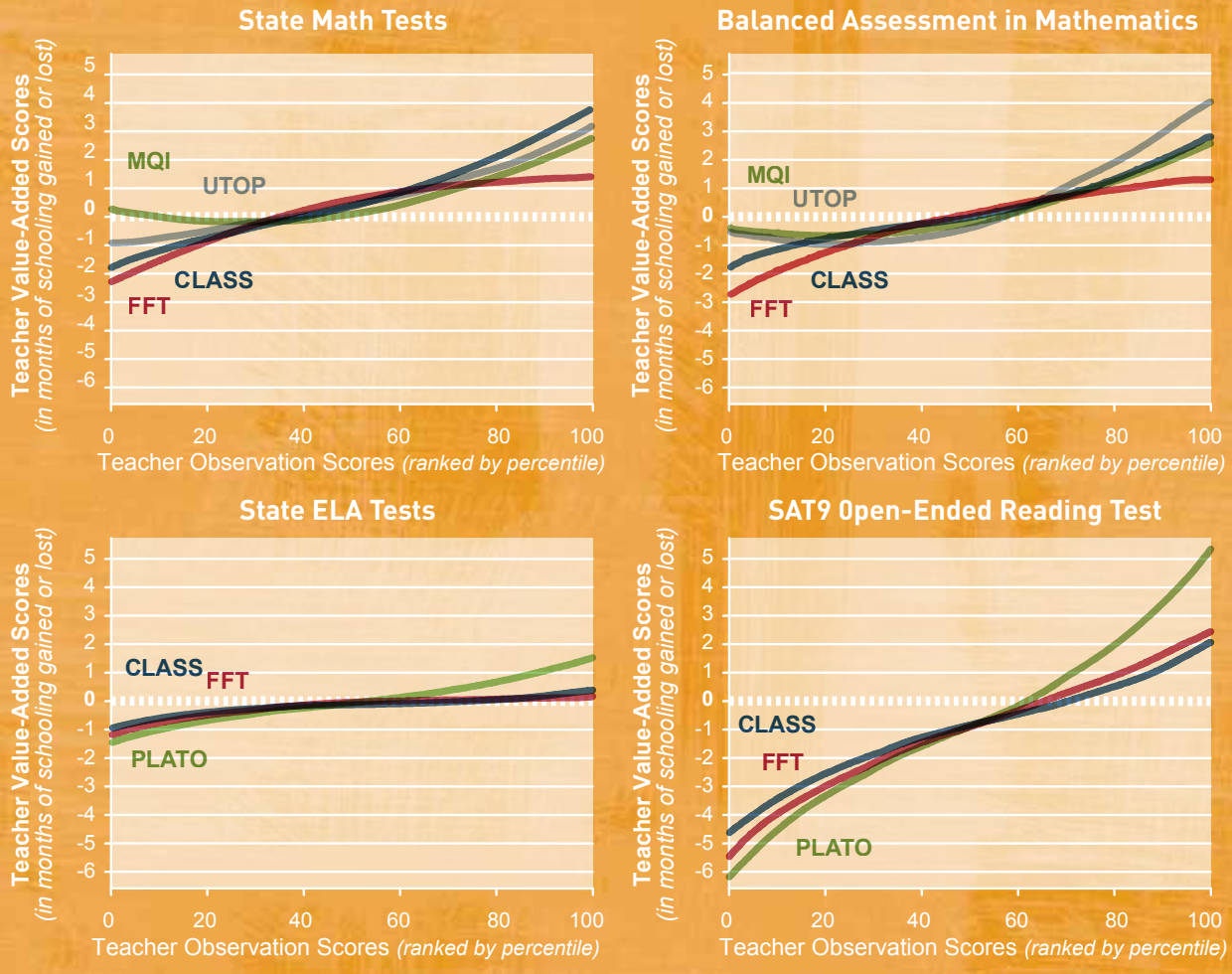
Observation Instruments and Gains on State Math and ELA Assessments

The teachers who demonstrated the types of practices emphasized in the classroom observation instruments had higher value-added scores than those who did not. **Figure 7,** on page 23, presents a graphical summary of the relationships. The position of the sloped lines indicates the average value-added scores (expressed in estimated months of schooling gained or lost relative to the average teacher) for teachers with different percentile rankings of observation scores.[13] As shown, as teachers' observation results increased, so did their value-added scores. This was true for all of the instruments.

Although statistically significant, many of these relationships did not indicate large differences in student learning based on observation scores alone. For example, the difference in student learning gains on state math tests between teachers in the top and bottom 25 percent of FFT scores amounted to approximately 2.7 months of schooling (assuming a nine-month school year). As evidenced in the bottom left panel in Figure 7, these differences were generally smaller in terms of value-added on state ELA tests than for math. For example, the estimated difference in student learning gains as measured on state ELA tests between the top and bottom 25 percent of teachers as ranked by their scores on FFT amounted to approximately 0.6 months of schooling.

Figure 7

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.

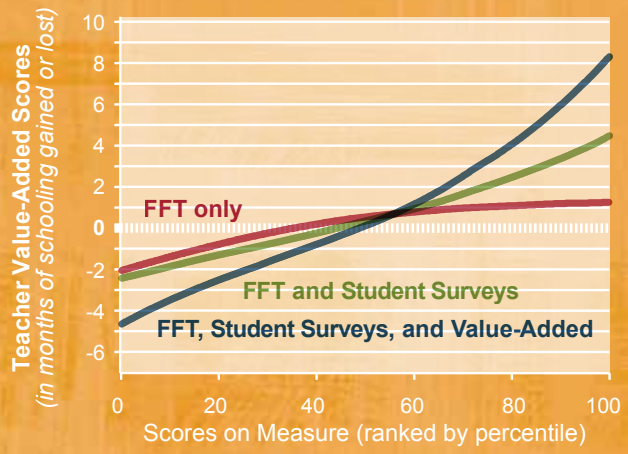
Observation Instruments and Gains on Supplemental Math and ELA Assessments

As mentioned earlier, we supplemented the state tests with more cognitively demanding assessments. When we tested the validity of classroom observations using the SAT9 OE reading test, we found stronger relationships than when the outcome was value-added on the state ELA tests. This is also shown in Figure 7, in which the steepness of the slope (showing the relationship between teachers' scores on each instrument and their students' gains as measured by the SAT9 OE) is far more pronounced than that for the state ELA test. In contrast to ELA, the value-added relationships with observation scores were similar across both types of math tests: state tests and the BAM tests. Researchers commonly find smaller differences between teachers on state ELA tests than on state math tests. Our findings, however, suggest that the reason may relate to the nature of state ELA tests, which often consist of multiple-choice questions of reading comprehension but don't ask students to write about their reading. After the early grades, however, many teachers have begun to incorporate writing into their ELA instruction. That may explain the greater relationship between observation scores and gains on the SAT9 OE. Regardless, the SAT9 OE seems to be more sensitive to teaching, as measured by classroom observations, than state ELA tests.

Figure 8

Combining Measures Added Predictive Power

Multiple Measures and Value-Added on State Math Test



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.

Combining Observations with Other Indicators

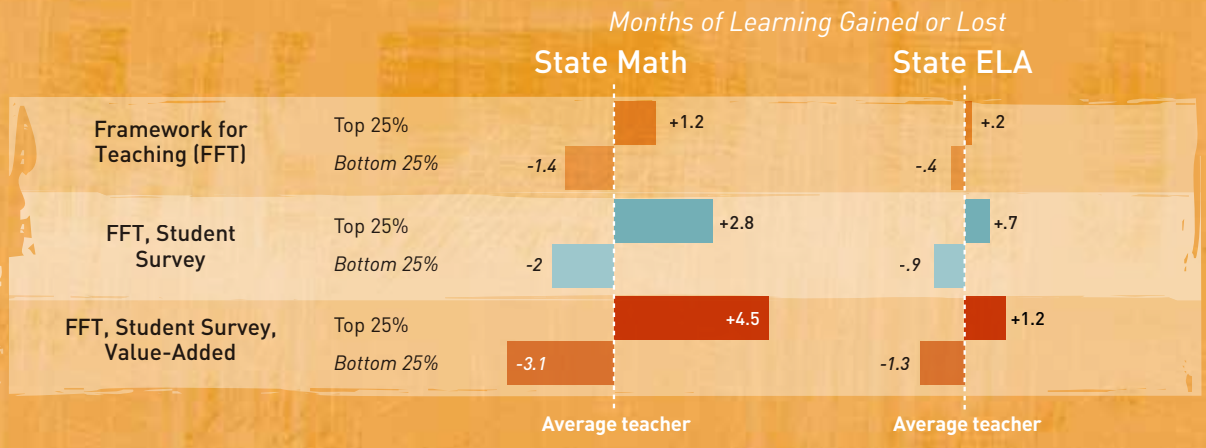
No measure is perfect. But better measures should allow for better decisions. To borrow a phrase from Lee Shulman, former head of the Carnegie Foundation for the Advancement of Teaching, the challenge school systems face is to assemble a "union of insufficient" measures that provide more information than they do individually and that are better than existing indicators. For this reason, we compared the relationship to student achievement outcomes of three different measures:

- Teachers' average scores on the classroom observation instruments alone;[14]
- A combination of teachers' observation scores and student feedback; and
- A combination of teachers' observation scores, student feedback, and value-added on state tests from another year or group of students.[15]

The pattern is clear: With each additional indicator, the relationship with student outcomes grew stronger. As shown in **Figure 8,** when going from FFT alone as an indicator to FFT plus student feedback, the relationship between learning gains on state math tests and teachers' scores grew stronger. And when FFT was combined with student feedback *and* value-added

Figure 9

Combining Observations with Other Measures Better Identified Effective Teaching



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.

gains, the combined measure's ability to predict a teacher's student achievement gains with another group of students grew even more.

Figure 9 presents these differences in terms of estimated months of schooling indicated by each measure. As shown, the difference in learning gains on state math tests between the top and bottom 25 percent of teachers increased from an estimated 2.6 months of learning to about 4.8 months when teachers were ranked on both FFT and the student survey. When value-added scores on state tests were added to the mix, the difference grew to 7.6 months—approaching the equivalent of an entire year of schooling. The same general pattern of increasingly strong associations with outcomes held when adding each of the five instruments to student feedback and value-added on state ELA tests. In other words, the difference in the magnitude of student learning gains between the high- and low-performing teachers, *as measured by the combination of three indicators,* was significantly greater than when classroom observations alone were the indicator. Combining the measures created the strongest indicator of effective teaching—one that was able to distinguish teaching practice that is associated with much greater learning gains.

As another yardstick to gauge the extent to which a combined measure predicted student learning differences, we also compared it to master's degrees and years of teaching experience—the two criteria most used for personnel decisions such as determining tenure, compensation, and the order in which teachers are considered for layoff during fiscal crises.

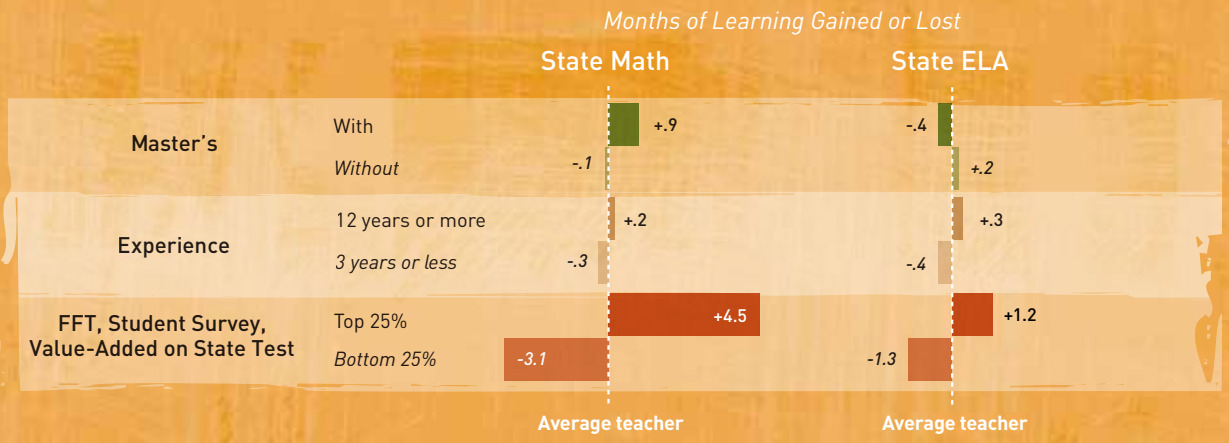
We found that among the teachers in our sample, the difference in learning gains on state math tests between those in the top and bottom 25 percent in terms of years of experience amounted to an estimated 0.5 months of schooling—less than one-tenth the difference indicated by the combined measure, including FFT, student feedback, and value-added. We found a 1.0-month difference in student learning, as measured by gains on state math tests, between teachers with and without master's degrees (see **Figure 10,** on page 26). The combined measure did a much better job than experience or master's degrees distinguishing among teachers with different achievement gains. (In fact, those with master's degrees on average had students who made *smaller* gains on state ELA tests than those without them.)

Relationships between Combined Measures and Outcomes other than State Tests

Stakeholders care about more than achievement on state tests. As a result, we looked at how teachers

Figure 10

Combined Measure 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.

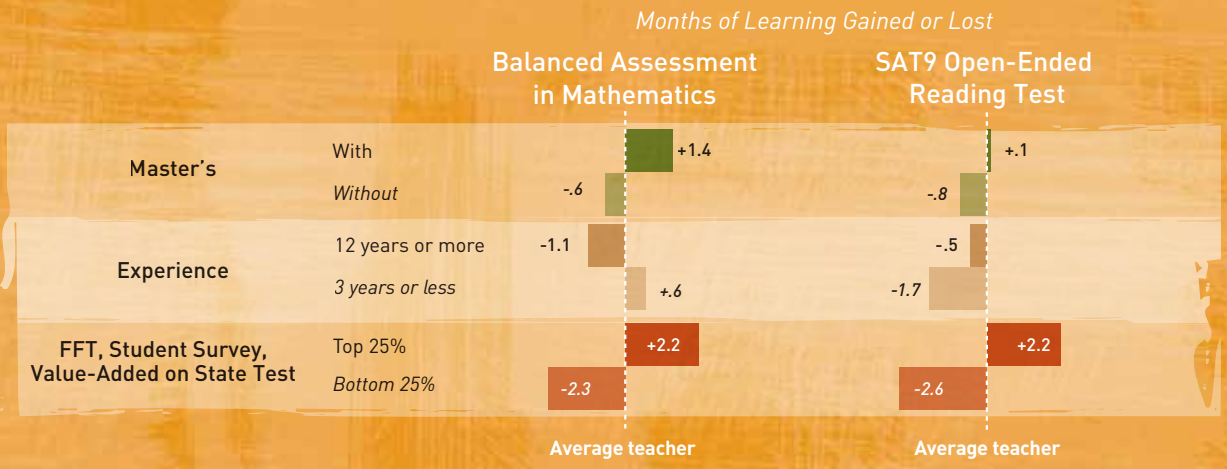
scoring well on the combined measure (including state test scores) performed on other student achievement outcomes. Through this project, we had a unique opportunity to administer student assessments that are more cognitively challenging and to look at self-reported student effort and student's positive attachment to school. When combining FFT scores with student feedback and student achievement gains on the state test, we found that the difference in estimated learning gains on BAM between the top and bottom 25 percent of teachers amounted to 4.5 months of schooling. Teachers who did well on a combined measure, which was based on state tests, tended to have students who did well on cognitively challenging assessments as well (see **Figure 11,** on page 27).

Of course, parents, teachers, and administrators also want students to enjoy school and feel engaged in their learning. For this reason, we also tested the relationship between a combined measure of teaching and survey items from our student feedback instrument indicating student effort (such as, "I have pushed myself hard to understand my lessons in this class") and students' positive emotional attachment to school (such as, "this class is a happy place for me to be"). In doing so, we again controlled for student characteristics, including prior achievement levels and peer achievement.

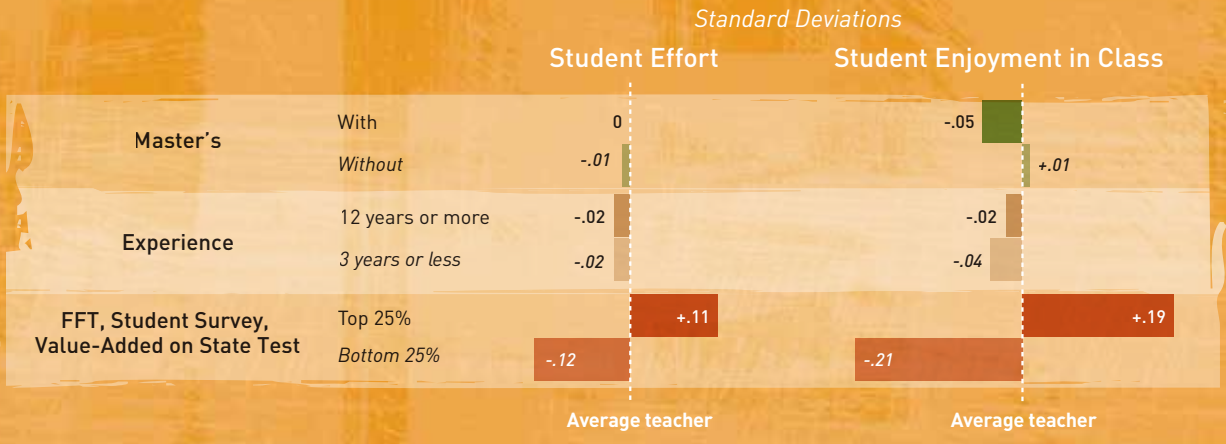
Because differences in outcomes such as reported effort and enjoyment cannot be equated to "months of learning," we gauged them in terms of standard deviations, a statistic for indicating differences within a distribution. But the comparison to master's degrees and years of experience makes the point: As shown in Figure 11, the combined measure identified teachers with bigger differences in student-reported effort and a positive emotional attachment to school.

Figure 11

Combined Measure Better Identified Effective Teaching on Supplemental Tests Than Master's or Experience



Combined Measure 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 ($p = 0.03$).

Challenges & Considerations for
Determining Alignment with Outcomes

Challenge

Teacher evaluation should improve teaching practice in ways that help teachers achieve greater success with their students. For that to happen, the measures must be related to student outcomes. But the need to do so raises a host of questions about indicators, outcomes, and processes.[16]

Considerations

How much alignment should be expected? Certainly the relationship should be positive, but the question of how much may depend on use. A modest relationship, as the MET project found for the instruments it tested, may give sufficient confidence for employing a measure as part of informal feedback. But when personnel decisions are at stake, the bar should rise. The MET project found the combined measure to be most aligned with a range of different student outcomes.

What about non-tested grades and subjects? A majority of teachers teach subjects and grades that are not included in state testing. But if measures of teaching, such as classroom observations or student feedback, are shown to be related to student outcomes in tested grades and subjects, it may be that they also are valid for non-tested ones.

Examples in Practice

Tennessee. The state department of education in Tennessee has created an information system to collect formal observation results from all districts in the state, allowing the agency to compare observation results with teachers' value-added scores and identify potential misalignment. In the cases of such misalignment, the state plans to work with districts to determine if the cause is grade inflation or some other factor.

Hillsborough Co., Fla. For each of its principals, the Hillsborough County Public Schools annually checks the alignment between the observation results of the schools' teachers and those same teachers' value-added scores.

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 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 is to periodically audit a representative sample of official observations, by having impartial observers perform additional observations. In our companion research 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 support teachers in improving their practice. But, by themselves, these measures are 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 not surprisingly, it is not as predictive of a teacher's achievement gains with other students as value-added. 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.

A Not-Final Word

Stay tuned. The findings discussed in this report represent but an update in the MET project's ongoing effort to support the work of states and districts engaged in reinventing the way teachers are evaluated and supported in their professional growth.

As a related effort, a separate soon-to-be-released report funded by the Bill & Melinda Gates Foundation will describe how leading systems are addressing the challenges associated with implementing quality classroom observations.

The next report from the MET project, anticipated by mid-2012, will use the project's extensive data set to deeply explore the implications of assigning different weights to different components of a system based on multiple measures of effective teaching—addressing a central question facing many state and district leaders. After that, we plan to release a report examining the extent to which student assignment may or may not play a role in measures of teacher effectiveness. This latter question is critical to address if measures are to be fair. To investigate the issue, we asked participating school leaders to create class rosters as they would normally and then to randomly assign teachers from among those who would normally teach them. The approach should remove any systematic bias in our measures resulting from the ways administrators assign students to teachers.

We often refer to the second of the documents just mentioned as the MET project's "Final Report." But the word "final" is a misnomer, as the MET project is making its data available for other researchers through partnership with the Inter-University Consortium for Political and Social Research at the University of Michigan. We expect this arrangement will produce an abundance of new analyses that further inform efforts to identify and develop effective teaching. Indeed, the MET project itself will continue to add to that understanding. Through a new stream of work—the MET Extension Project—we are returning to the classrooms of some of the MET project volunteers with the goal of videotaping many more lessons using a new generation of cameras so we can continue analysis and produce a video library of practice for use in teacher professional development. These videos also will be incorporated into tools we currently are developing that will enable states and districts to certify observers and to validate their own observation instruments.

In the meantime, we hope an underlying message in our work is not missed. The MET project is investigating measures of teaching not merely to produce findings but also to model what it means to ask the right questions. States and districts should themselves ask: How accurate are our observers? How reliable are our observation procedures? How aligned are our measures to student outcomes? The imperatives of quality feedback and improvement demand it.

Participant Perspectives

Although very much in the spirit of professional development, the MET project is ultimately a research project. Nonetheless, participants frequently told us they have grown professionally as a result of their involvement. MET project volunteer teachers had the opportunity to see themselves teaching when they reviewed their lesson videos prior to submitting them. Lesson raters got the chance to learn the observation instruments in depth and also analyze many hours of instruction. Below is a sampling of comments we received from teachers and raters, which we think speak to the value of classroom observations.

From Teachers

"The videotaping is what really drew me in, I wanted to see not only what I'm doing but what my students are doing. I thought I had a pretty good grasp of what I was doing as a teacher, but it is eye opening … . I honestly felt like this is one of the best things that I have ever done to help me grow professionally. And my kids really benefited from it, so it was very exciting."

"With the videos, you get to see yourself in a different way. Actually you never really get to see yourself until you see a video of yourself. I changed immediately certain things that I did that I didn't like."

"I realized I learned more about who I actually was as a teacher by looking at the video. I learned of the things that I do that I think that I'm great at I was not so great at after all."

"Even the things I did well, I thought, OK that's pretty good, why do I do that, and where could I put that to make it go farther. So it was a two-way road, seeing what you do well, and seeing the things that have become habits that you don't even think about anymore."

From Raters

"Being a rater has been a positive experience for me. I find myself 'watching' my own teaching more and am more aware of the things I should be doing more of in my classroom."

"I have to say, that as a teacher, even the training has helped me refine my work in the classroom. How wonderful!"

"Being a rater has helped me become a much better teacher and evaluator."

"I have loved observing teachers, [being a rater for the MET project has helped me in] reflecting on my own teaching and that of the teachers teaching in my school."

Endnotes

1 Much help and advice came from Jeff Archer, Sarah Buhayar, Steve Cantrell, Todd Kawakito, Kerri Kerr, and David Parker. KSA-Plus Communications provided editorial and design assistance.

2 These quality assurance strategies depend on accurate links between student and teacher data, without which trust in the system will be compromised. Teachers should have an opportunity to verify the rosters of students for whom they are responsible.

3 For more information on the Tripod student survey instrument used in this analysis, see *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.*

4 This number is a subset of the approximately 2,700 teachers from whom the MET project collected video. The sample does not include 9th grade teachers, and it excludes most teachers who were not assigned classes through a randomized process as part of an investigation into the effects of student assignment on teacher effectiveness measures—the focus of an upcoming report. See Appendix Table 1 in the research report for more detail.

5 Other parts of FFT allow that use of rote questions may be appropriate at times, as for reviewing with students, but not as a way to deepen students' understanding.

6 Different instruments call these "dimensions," "components," and "elements." For the sake of consistency in discussing multiple instruments, we use "competencies."

7 For evidence that observation-based teacher coaching improves student achievement, see: 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. The coaching model described, based on CLASS, is more fully explained for practitioners in the paper "Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training," by Robert Pianta, from the Center for American Progress (November 2011).

8 One of our partners, the National Board for Professional Teaching Standards, has provided data for those applying for certification from the MET project districts. The MET project also is investigating a sixth observation instrument, Quality Science Teaching (QST), developed by Raymond Pecheone and Susan E. Schultz at Stanford University. QST focuses on high school instruction and so is not included in the initial analysis in this report on results from grades 4–8. Results from both of these will be included in our final report in mid-2012.

9 UTOP training, managed by the National Math and Science Initiative (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, the 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. The UTOP results also are not directly comparable to the results for the other instruments because they are based on different samples of teachers. For more on how the UTOP sample and scoring differed, see the companion research report.

10 By "impartial" we mean someone who is not biased by the same familiarity with the teacher as the person who gave the original score.

11 For more information on BAM and SAT9 OE, see *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.*

12 While a worthwhile precaution, the use of different students for the instructional measures and outcomes is not a guarantee against bias. For instance, if a teacher is always assigned students with the same unmeasured trait, this strategy will not reduce the bias. In the end, the only way to resolve this question is to randomly assign one group of students to teachers and to ensure that the outcomes and instructional measures are captured with different groups of students present. In our final report in mid-2012, we will focus on just such a comparison.

13 The months of schooling estimate is based on the assumption that nine months equals 0.25 standard deviation difference in achievement. For information on differences in terms of standard deviations and how these differences relate to correlation calculations, see the companion research report.

14 The figure uses observation scores from one section and relates them to a teacher's value-added with another course section. As described in the companion research report, we obtained similar results using observation scores from the 2009–10 school year and relating them to value-added for the same teachers in 2008–09.

15 These comparisons are based on equal weights of each component in the combined measure, so that classroom observations plus student feedback are weighted 50/50, and observations plus feedback plus value-added are weighted 33/33/33. For more on this topic, see the companion technical report.

16 The Brookings Institution has published a report suggesting ways a state or district could use its data to assess the alignment of its teacher evaluation system: *Passing Muster: Evaluating Teacher Evaluation Systems.*

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.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org