MET project

POLICY AND PRACTICE SUMMARY

Gathering Feedback for Teaching

Combining High-Quality Observations with Student Surveys and Achievement Gains

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.

In a previous paper, we reported that confidential student surveys about students' classroom experiences can provide reliable and meaningful feedback on teaching practice. 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. 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.

BILL & MELINDA GATES foundation

Major 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.

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.

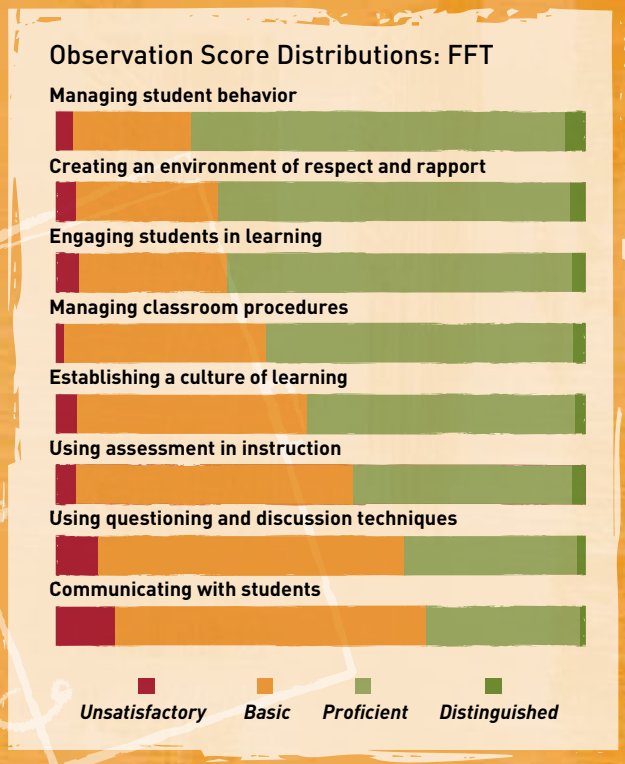
Therefore, we emphasize the following six minimum requirements for high-quality classroom observations:

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.

Defining & Diagnosing Teaching Practice

This chart shows the distribution of scores given to lesson videos of MET project teachers on eight competencies from the Framework for Teaching. Results from the other four observation instruments studied by the project show similar patterns: most scores were in the mid-range; and more high scores were given for classroom management than for complex aspects of instruction such as questioning techniques.



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.

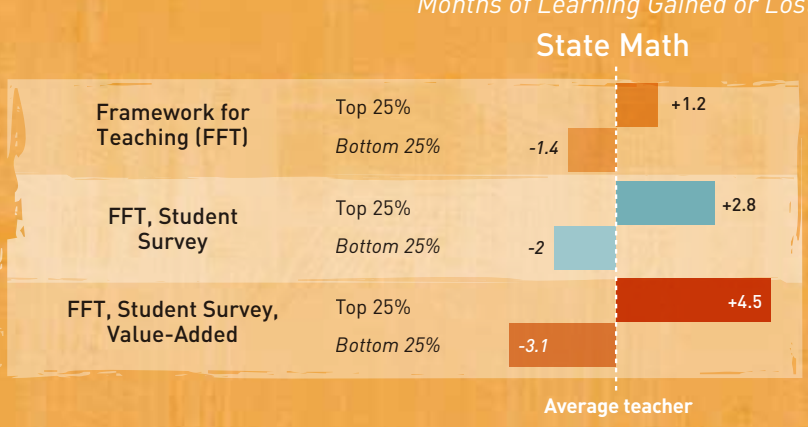
5. **Combine observations with student achievement gains and student feedback.** The combination of classroom observations, student feedback, and student achievement carries three advantages over any measure by itself: (a) it increases the ability to predict if a teacher will have positive student outcomes 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*.

Combining Observations with Other Measures

Better Identified Effective Teaching

Based on observation results alone, students who had the top 25% of teachers gained 1.2 months of learning on state math tests (relative to the average teacher), while students who had the bottom 25% lost 1.4 months — a gap of 2.6 months. Combined measures, however, were better able to distinguish among teachers with different student achievement gains. Using a measure that included observation, student survey, and value-added results, students taught by top 25% of teachers gained about 4.5 months of schooling, while those taught by teachers in the bottom 25% lost 3.1 months — a gap of 7.6 months.



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.

Stay tuned.

The findings discussed here 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. Coming up: a report that explores the implications of assigning different weights to different components of a system based on multiple measures of effective teaching.

To download the full brief and research paper, plus other MET reports, go to www.metproject.org.

The MET project is a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org