Gathering Feedback for Teaching

Combining High-Quality Observations with Student Surveys and Achievement Gains

Bill & Melinda Gates Foundation
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.

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), and
- **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.

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.