The MET Project

In fall 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to test new approaches to recognizing effective teaching. The project’s goal is to help build fair and reliable systems for teacher observation and feedback to help teachers improve and administrators make better personnel decisions. With funding from the foundation, the data collection and analysis are being led by researchers from academic institutions, nonprofit organizations, and several private firms and are being carried out in seven urban school districts.

Research Partners

Our lead research partners include:
- Mark Atkinson, Teacshcape
- Nancy Caldwell, Westat
- Ron Ferguson, Harvard University
- Drew Gitomer, Educational Testing Service
- Eric Hirsch, New Teacher Center
- Dan McCaffrey, RAND
- Roy Pea, Stanford University
- Geoffrey Phelps, Educational Testing Service
- Rob Ramseid, Cambridge Education
- Doug Staiger, Dartmouth College
- Catherine McClellan, Educational Testing Service
- Denis Newman, Empirical Education
- Raymond Pecheone, Stanford University
- Robert Pianta, University of Virginia
- Morgan Polikoff, University of Southern California
- Steve Raudenbush, University of Chicago
- John Winn, National Math and Science Initiative

Key contributors include:
- Joan Aucott, National Board for Professional Teaching Standards
- Charlotte Danielson, The Danielson Group
- Dan Goldhaber, University of Washington
- Pam Grossman, Stanford University
- Bridget Hamre, University of Virginia
- Heather Hill, Harvard University
- Sabrina Laine, American Institutes for Research

For four decades, education research has confirmed what many parents know: A child’s learning depends on the talent and skills of the person leading his or her classroom. As much as parents worry about their local school, most eventually learn that their child’s teacher in that school matters even more.

Yet most school systems ignore differences among individual teachers. Information about teaching effectiveness is neither collected nor shared. The costs of this neglect are enormous. Novice teachers’ skills plateau too early without feedback they need to improve. Likewise, there are too few opportunities for experienced teachers to share their practice and strengthen the profession. Finally, principals are forced to make the most important decision they ask of them—granting tenure to beginning teachers early in their careers—with little objective information to guide them.

If we say “teachers matter” and the research clearly says they do, why do we pay so little attention to the work teachers do in the classroom? If teachers are producing dramatically different results, why don’t we provide them with feedback and trust them to respond?

The MET Project

In fall 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to test new approaches to recognizing effective teaching. Our goal is to help build fair and reliable systems for teacher observation and feedback to help teachers improve and administrators make better personnel decisions.

The Effective Teaching Pathway

- Multiple measures of effectiveness
- Accurate teacher evaluation
- More meaningful tenure
- Differentiated pay based on effectiveness
- More effective teachers
- Strategic placement of teachers
- Targeted PD and other teacher supports
- BETTER STUDENT OUTCOMES
classrooms and by asking students. If so, such measures would be useful for both developing teachers and staffing schools more effectively.

With funding from the foundation, the data collection and analysis is being led by researchers from academic institutions, nonprofit organizations, and several private firms. However, the hardest work is being done by the 3,000 teacher-volunteers, working in seven urban school districts (New York City, Charlotte-Mecklenburg, Hillsborough County in Florida, Memphis, Dallas, Denver, and Pittsburgh), who have agreed to open their classrooms.

Although the project is ongoing (the final report will not be released until winter 2011-12), we are reporting our findings so that such measures would be useful for schools more effectively.

With funding from the foundation, the data collection and analysis are being led by researchers from academic institutions, nonprofit organizations, and several private firms. However, the hardest work is being done by the 3,000 teacher-volunteers, working in seven urban school districts (New York City, Charlotte-Mecklenburg, Hillsborough County in Florida, Memphis, Dallas, Denver, and Pittsburgh), who have agreed to open their classrooms.

Although the project is ongoing (the final report will not be released until winter 2011-12), we are reporting our findings so that such measures would be useful for schools more effectively.

Our Analysis

As a school leader staffs a school each year, he or she must ask, “What does each teacher’s past performance say about his or her ability to help students learn?” and “What are his or her specific strengths and weaknesses?” Every artifact of a teacher’s practice—whether student surveys about a teacher’s effectiveness, direct classroom observations, or [in an increasing number of school districts] the achievement gains of recent or past students—is potentially useful in identifying a teacher’s strengths and weaknesses and prospects of success with future students. Effective leaders can use such data to guide a teacher’s development.

Our analysis plan mimics the school leader’s questions. We ask, “How well do various aspects of a teacher’s performance in one course section or in one academic year help predict student achievement gains in that teacher’s classroom during another academic year or in another course section?” In this preliminary report, we measure student achievement gains using two different tests in each subject, the state standardized test and an additional, more rigorous test. For now, we test the value of evidence of effectiveness from one class in anticipating student achievement gains in another class taught by the same teacher. To do that, we use two analogous thought experiments:

- First, focusing on the subset of teachers for whom we have measures from more than one classroom of students during 2009-10, we ask whether the measures of practice from one class predict the teacher’s contribution to student learning gains in another class.
- Second, focusing on those teachers for whom we have student assessment data from a prior year (2008-09), we test whether measures of classroom practice in 2009-10 are related to the teacher’s contribution to student learning gains in another school year.

---

1 Pittsburgh served as our pilot district, an important role, but no data from this district will be analyzed.
2 The Tripod survey, which we used, was developed over the past decade by Dr. Ron Ferguson from Harvard in collaboration with Cambridge Education.
3 The results from Memphis have been delayed because of a new state test in Tennessee last spring. Moreover, we are still organizing the data for the high school students from the other districts.
Early Findings

Although the accompanying technical report provides more details on our analysis and initial results, we have four general findings to report:

First, in every grade and subject we studied, a teacher’s past success in raising student achievement on state tests (that is, his or her value-added) is one of the strongest predictors of his or her ability to do so again. When applied to teaching, the term value-added refers to statistical efforts to isolate the impact of a teacher on his or her students’ achievement by adjusting for each student’s starting point coming into the class. Each student’s performance at the end of the year is then compared to that of similar students elsewhere (with similar prior test scores, similar demographics, etc.). When a teacher’s students outperform his or her peers whose students have similar prior achievement, characteristics, and classrooms, it constitutes positive student growth or value-added. In this analysis, we also adjusted for the mean characteristics of the other students in the class, since one’s peers also can have an influence on one’s learning. Conversely, when a teacher’s students perform worse than his or her peers whose students have similar starting points and similar classmates, it constitutes negative growth or value-added.

A teacher’s history of positive (or negative) value-added is among the strongest predictors of his or her students’ achievement growth in other classes and academic years. Value-added methods have been criticized as being too imprecise, since they depend on the performance of a limited number of students in each classroom. Indeed, we do find that a teacher’s value-added fluctuates from year to year and from class to class, as succeeding cohorts of students move through his or her classrooms. However, our analysis shows that volatility is not so large as to undercut the usefulness of value-added as an indicator of future performance.

Second, the teachers with the highest value-added scores on state tests also tend to help students understand math concepts or demonstrate reading comprehension through writing.

Many have speculated that teachers who are eager to improve. Students’ perceptions have two other welcome characteristics: They provide a potentially important measure that can be used in nontested grades and subjects. In addition, the information received by the teacher is more specific and actionable than value-added scores or test results alone.

Fourth, valid feedback need not be limited to test scores alone. By combining different sources of data, it is possible to provide diagnostic, targeted feedback to teachers who are eager to improve.
The public discussion usually portrays only two options: the status quo (where there is no meaningful feedback for teachers) and a seemingly extreme world in which tests scores alone determine a teacher’s fate. Our results suggest that’s a false choice. It is possible to combine measures from different sources to get a more complete picture of teaching practice. The measures should allow a school leader to both discern a teacher’s ability to produce results and offer specific diagnostic feedback. Value-added scores alone, while important, do not recommend specific ways for teachers to improve.

Ultimately, we will be adding data from classroom observations and a new teacher assessment to the mix of measures we are testing. However, our initial analyses suggest that the combined measures help identify effective and ineffective teaching. For example, we used evidence of a teacher’s performance (as measured by value-added and student perceptions) in one class to infer which teachers were more and less effective. We then assessed the impact of these teachers on learning gains for a different group of students. As shown by the “State Math Test” bar in the graphic above, students of math teachers whose value-added scores and student perceptions placed them in the bottom 25 percent gained the equivalent of only six and a half months of learning during a nine-month school year as compared to the average teacher. The number of months of schooling applies to a nine-month school year, using a .25 standard deviation per year conversion factor.

The MET project. We still need to complete the analysis of 13,000 classroom lessons observed during the 2009–10 school year and the fresh set of lessons from the current school year. We will test a new measure that extends and refines the concept of pedagogical content knowledge for teachers, or what a teacher knows about how to teach a particular subject. These findings could have significant implications, not only for measuring effective teaching but for the training and development of teachers as well. In late spring 2011, we will issue a more complete report from year one that includes findings from the classroom observation protocols. Late summer of 2011 researchers from RAND will combine data from each of the MET project measures to form a “composite indicator” of effective teaching. Researchers from RAND will analyze different approaches to weighting each measure (student achievement on state and supplemental tests, classroom observations, teacher knowledge, student perceptions) when forming an overall assessment of a teacher’s effectiveness. Finally, early in 2012, we will report whether those teachers whose performance was rated most highly during the 2009–10 school year actually produced larger student achievement gains than their colleagues during the 2010–11 school year.

### Students with Most Effective Teachers Learn More in a School Year

<table>
<thead>
<tr>
<th>Quarter of Teachers with Least Evidence of Effectiveness</th>
<th>Quarter of Teachers with Most Evidence of Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.7 months</td>
<td>+4.8 months</td>
</tr>
<tr>
<td>-3.2 months</td>
<td>+2.9 months</td>
</tr>
<tr>
<td>-1.4 months</td>
<td>+1.4 months</td>
</tr>
<tr>
<td>-5.8 months</td>
<td>+5.0 months</td>
</tr>
</tbody>
</table>

**State Math Test**
-2.7 months
-3.2 months
-1.4 months
-5.8 months
+4.8 months
+2.9 months
+1.4 months
+5.0 months

**Balanced Assessment of Mathematics**
-1.4 months

**State ELA Test**
+1.4 months

**SAT9/Open-Ended Reading**
-5.8 months

Months of learning gain are calculated based on the difference in value-added gains between the top and bottom quartile of teachers compared to the average teacher. The number of months of schooling applies to a nine-month school year, using a .25 standard deviation per year conversion factor.

The measures help identify effective and ineffective teaching. For example, we used evidence of a teacher’s performance (as measured by value-added and student perceptions) in one class to infer which teachers were more and less effective. We then assessed the impact of these teachers on learning gains for a different group of students. As shown by the “State Math Test” bar in the graphic above, students of math teachers whose value-added scores and student perceptions placed them in the bottom 25 percent gained the equivalent of only six and a half months of learning during a nine-month school year as compared to the average teacher. The number of months of schooling applies to a nine-month school year, using a .25 standard deviation per year conversion factor.

### Measures of Effective Teaching (MET) Project

<table>
<thead>
<tr>
<th>Implementation Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fall 2010</strong></td>
</tr>
<tr>
<td>Preliminary results from year one data collection: student perception survey and associated student achievement gains on multiple assessments</td>
</tr>
<tr>
<td><strong>Winter 2011</strong></td>
</tr>
<tr>
<td>Preparing systems for multiple measures of teacher evaluation: using digital video, training observers, and meeting data requirements</td>
</tr>
<tr>
<td><strong>Spring 2011</strong></td>
</tr>
<tr>
<td>Expanded results from year one (predictors of teaching effectiveness): sample of observed lessons and correlation with value-added assessments</td>
</tr>
<tr>
<td><strong>Summer 2011</strong></td>
</tr>
<tr>
<td>Technical report on composite measure of effective teaching</td>
</tr>
<tr>
<td><strong>Winter 2011–12</strong></td>
</tr>
<tr>
<td>Final results</td>
</tr>
</tbody>
</table>

---

First we sorted teachers based on student perception surveys and value-added on the state math assessment. Then we sorted teachers into quartiles. The percentage of students agreeing above the mean for the top and the bottom quartile teachers.
Conclusion

Reinventing the way we develop and evaluate teachers will require a thorough culture change in our schools. No longer should teachers expect to close the door to their classrooms and “go it alone.” The quality of instruction is a public good, and improvement will require a collective commitment to excellence in every classroom.

Teachers will need to open up their practice for review and constructive critique—because that’s what excellence requires.

There are some obvious places to start now:

- working with teachers to develop accurate lists of the students in their care, so that value-added data are as accurate as possible
- using confidential surveys to collect student feedback on specific aspects of a teacher’s practice, including those in nontested grades and subjects
- retraining those who do classroom observations to provide more meaningful feedback

While we still have much to learn in the future, we can now confidently encourage states and districts to regularly check that the collection of measures they assemble allows them to explain the variation in student achievement gains among teachers. Just as we have done in the accompanying report, they should confirm that the measures as implemented continue to point in the same direction. Even a great classroom observation tool can be implemented poorly (if principals are poorly trained or if they are unwilling to provide honest feedback). Even a great instrument for collecting student feedback can be distorted (if students do not take it seriously or if students do not trust that their answers will be kept confidential). The best way to ensure that the evaluation system is providing valid and reliable feedback to teachers is to regularly verify that—on average—those who shine in their evaluations are producing larger student achievement gains.

Since we are just starting, we need to be humble about what we know and do not know. However, we should take heart in the fact that the solutions to our educational challenges are implemented every day by those teachers who regularly generate impressive results. We just need to assemble the evidence on student achievement, ask students to help by providing their own confidential feedback, and refine our approach to classroom observation—to find those teachers who truly excel, support them, and develop others to generate similar results. The MET project is an important first step.