MET
project

Working with Teachers to Develop Fair and Reliable Measures of

Effective Teaching

BILL & MELINDA
GATES *foundation*

Bill & Melinda Gates Foundation

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—especially those with the fewest resources—have access to the opportunities they need to succeed in school and life. Based in Seattle, Washington, the foundation is led by CEO Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

For more information on the U.S. Program, which works primarily to improve high school and postsecondary education, please visit www.gatesfoundation.org.

June 2010

Introduction

In fall 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to develop and test multiple measures of teacher effectiveness. The goal of the MET project is to improve the quality of information about teaching effectiveness available to education professionals within states and districts—information that will help them build fair and reliable systems for teacher observation that can be used for a variety of purposes, including feedback, development, and continuous improvement.
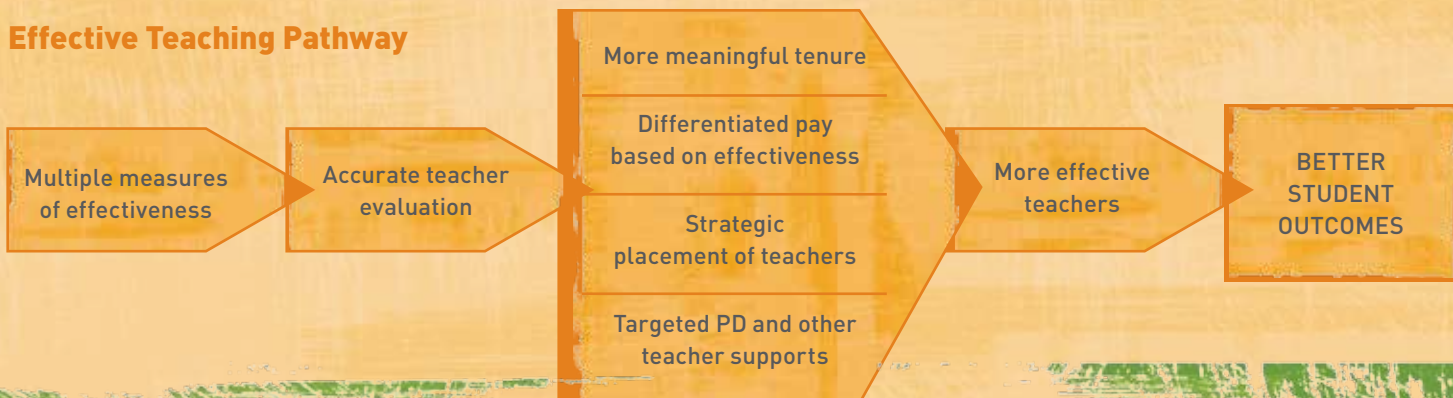
This information will include video-taped classroom observations, student surveys, tests of teachers' pedagogical content knowledge, and analyses of student assessment data to examine achievement gains over time. A close analysis of these indicators will help establish which teaching practices, skills, and knowledge positively impact student learning.

A teacher's effectiveness has more impact on student learning than any other factor controlled by school systems, including class size, school size, and the quality of after-school programs—or even which school a student is attending.[1] In a study of Los Angeles schools, for instance, the difference between the performance of a student assigned to a top-quartile teacher and a student assigned to a bottom-quartile teacher averaged 10 percentile points on a standardized math test.[2] Researchers studying high schools in North Carolina found that having a class with a strong teacher produced results 14 times greater than having a class with five fewer students.[3]

Dramatically improving education means ensuring that every student has an effective teacher in every classroom every school year. Better information about teacher effectiveness could be an extraordinarily valuable tool for achieving this goal. If the average classroom of tomorrow is as productive as the top quarter of our classrooms today, the United States could close the gap in achievement with higher-performing countries, such as Japan, within two years.[4]

Effective Teaching Pathway

Multiple measures of effectiveness → Accurate teacher evaluation → More meaningful tenure / Differentiated pay based on effectiveness / Strategic placement of teachers / Targeted PD and other teacher supports → More effective teachers → BETTER STUDENT OUTCOMES

The Problem

Current teacher evaluation systems are not providing the information needed to close the achievement gap. Despite 40 years of research pointing to huge differences in student achievement gains across teachers, most school districts and state governments cannot pinpoint what makes a teacher effective or identify their most and least effective teachers.

Although there is growing consensus that effective teaching is the key to large-scale school reform, there is no agreement among education stakeholders about how to identify and measure effective teaching. Almost everywhere, teacher evaluation does not provide meaningful feedback to help teachers improve. Nor does it provide supervisors with the objective data they need to make informed assessments of teachers' strengths and weaknesses.

Rather, most evaluation is a perfunctory exercise based largely on characteristics unrelated to student achievement. The 2009 New Teacher Project study *The Widget Effect,* for example, found that for evaluation systems with two ratings, "satisfactory" and "unsatisfactory," 99 percent of teachers earned a
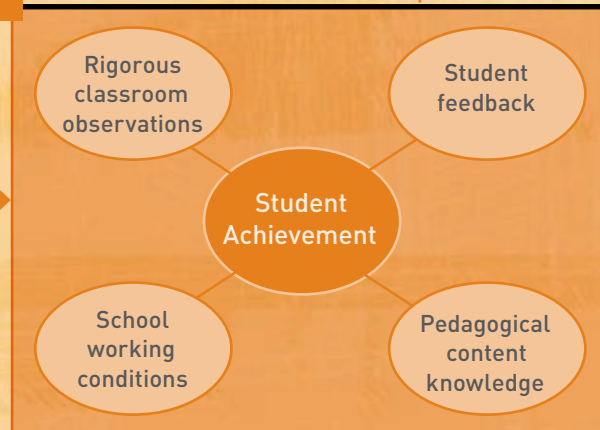
Proposed Teacher Evaluation and Development Criteria

Today

- High-level principal input
- Seniority
- Degrees earned
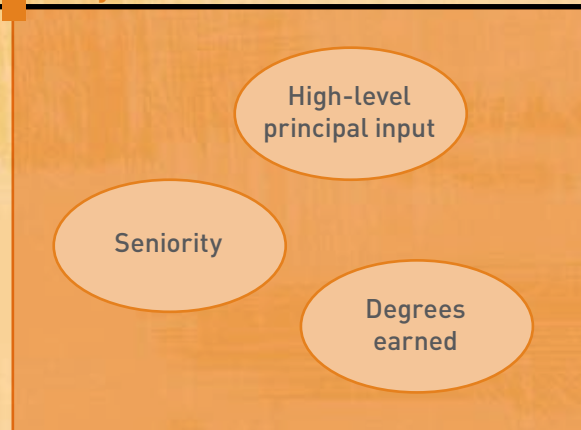
Basic: Principal observation and teacher "qualifications" determine rating

Future Scenario (to be developed and tested)

- Rigorous classroom observations
- Student feedback
- Student Achievement
- School working conditions
- Pedagogical content knowledge

Robust: Multiple inputs anchored in student achievement determine effectiveness

"By identifying what methods work well in a classroom, we have the potential to improve outcomes for many more of our students."

*—**Joel I. Klein**, Chancellor, New York City Department of Education*

satisfactory. In evaluation systems with more than two ratings, 94 percent of teachers received one of the top two ratings and less than 1 percent were rated unsatisfactory.[5]

Even in the rare instances in which evaluation systems are directly linked to student achievement, measures of teaching generally rely on student test scores as the exclusive proxy for effectiveness. They rarely take into account the full range of what teachers do or the context in which they teach.

The absence of good information about teacher effectiveness limits the ability of district and school administrators to make informed decisions about teacher recruitment, evaluation, development, placement, tenure, compensation, and retention. Students with the greatest needs clearly require the most effective teachers. But far too few systems know which teachers to place with the neediest students, and some systems even prohibit principals from making placement or hiring decisions based on teacher effectiveness.

In the absence of useful feedback, most teachers' performance plateau by their third or fourth year on the job.[6] Everyone loses as a result. Students are shortchanged by teachers whose careers have effectively stalled; many of them disengage. With few incentives, insufficient guidance, and a lack of professional learning and support, many promising teachers leave the profession for other occupations and industries, while good teachers are demoralized by ineffective colleagues. And the nation compromises its future productivity and competitiveness by not educating all young people to their full potential.

The Response

The MET project, supported by the Bill & Melinda Gates Foundation, will provide a new knowledge base for practitioners and policymakers who are trying to strengthen the teaching profession. By incorporating student achievement gains, direct student feedback, videotaped classroom observations, and new assessments of teachers' pedagogical content knowledge, the MET project will share insights and develop new tools that will make evaluation a more valuable professional opportunity for teachers, while allowing districts and states to develop more meaningful and effective processes and policies.

The MET project is led by more than a dozen organizations: academic institutions (Dartmouth College, Harvard University, Stanford University, University of Chicago, University of Michigan, University of Virginia, and University of Washington), nonprofit organizations (Educational Testing Service, RAND Corporation, and the New Teacher Center), and several for-profit education consultants (Cambridge Education, Teachscape, and Westat). In addition, the National Board for Professional Teaching Standards and Teach For America are supporting the project and have encouraged their members to participate. The American Federation of Teachers and the National Education Association have been involved in discussions about the MET project and support the research.

Research Design Considerations

For a variety of measures of effective teaching to be used, they must be based on aspects of teaching that excellent teachers recognize as characteristic of their practice; if the measures are unrecognizable to thoughtful practitioners, they will not be adopted. Similarly, for measures of effective teaching to be effective, they must pinpoint aspects of teaching that improve student learning; if the measures are unrelated to student learning, they will have no impact.

The MET project is based on two simple premises: First, a teacher's evaluation should depend to a significant extent on his/her students' achievement gains; second, any additional components of the evaluation (e.g., classroom

> *"Educators know all too well that one-dimensional indicators such as test scores can't begin to capture the complexities of effective teaching and learning. This study promises to look at the bigger picture, and we view it as an important opportunity to be proactive about our profession."*
>
> **—Michael Mulgrew,** *President, United Federation of Teachers, New York City*

observations) should be valid predictors of student achievement gains.[7]

While committed to the use of student assessments to help measure teacher effectiveness, the MET project recognizes that teacher evaluation cannot depend on student test gains alone.

- Not all subjects and grades currently have mandated tests. As a result, if teacher effectiveness measures were limited to a score based on teachers' contributions to student performance on standardized tests, the feedback would exclude the majority of teachers— all of whom have an important role in student learning. This issue could be resolved by additional tests, but tests are resource- and time-intensive (for both students and teachers) and are highly variable in quality.

- Value-added measures, which determine a teacher's unique contribution to each student's performance,

offer fair comparisons among teachers within a system, but they do not and cannot help teachers understand why one teacher is more successful than another. Teachers with the highest and lowest value-added scores are both left to speculate about what they did to merit their scores. More important, the scores do not suggest what a teacher would have to change to improve his/her effectiveness in the classroom.

- For some teachers, particularly those early in their careers, consequential performance judgments would be made based on the test performance of relatively few students. Though this concern diminishes over time, multiple measures could allow more

accurate judgments earlier in a teacher's career when they could have a significant impact on a teacher's professional growth.

For these reasons, non-test-based measures also are needed, but they are not all equally useful and should be treated accordingly. Certain measures should carry more weight than others. Relative weights should depend on the measures' demonstrated track record of improving student achievement gains.

Data Collection: Sites, Partners, and Measures

To help identify the best mix of teacher effectiveness measures, more than 3,000 teacher volunteers are participating in the MET project across six predominantly urban school districts: Charlotte-Mecklenburg Schools, Dallas Independent School District, Denver Public Schools, Hillsborough County Public Schools, Memphis City Schools, and the New York City Department of Education. Participants teach math and
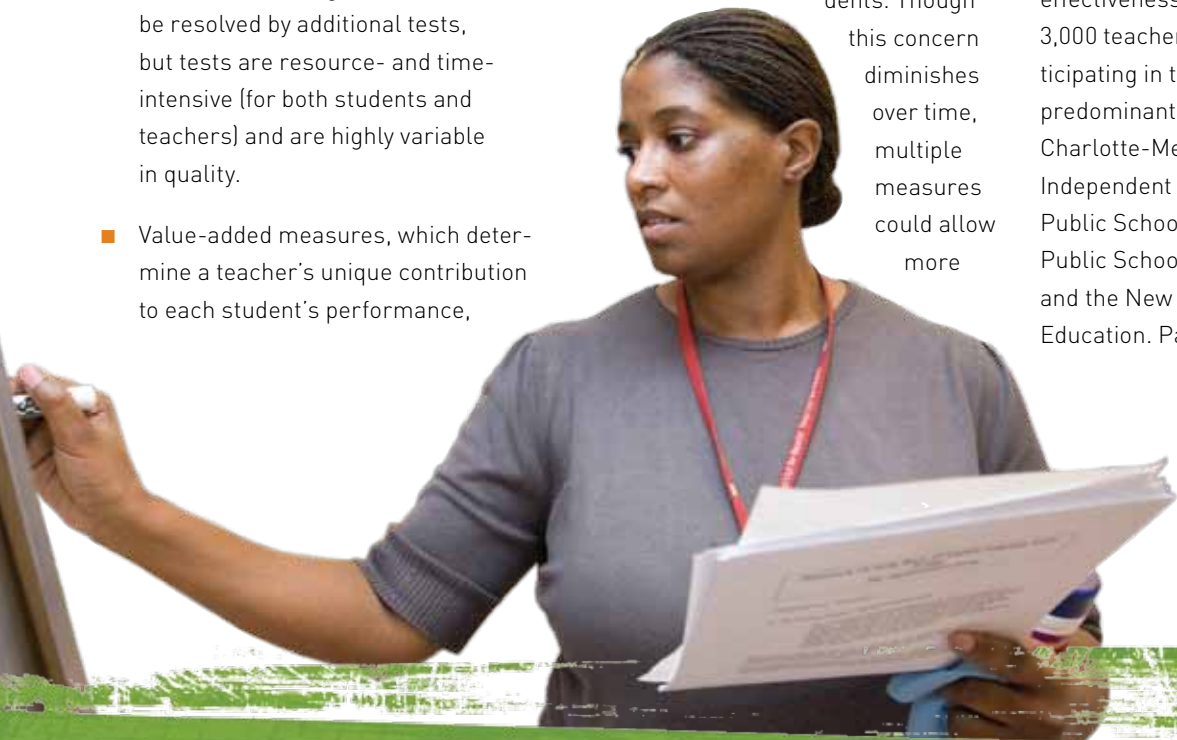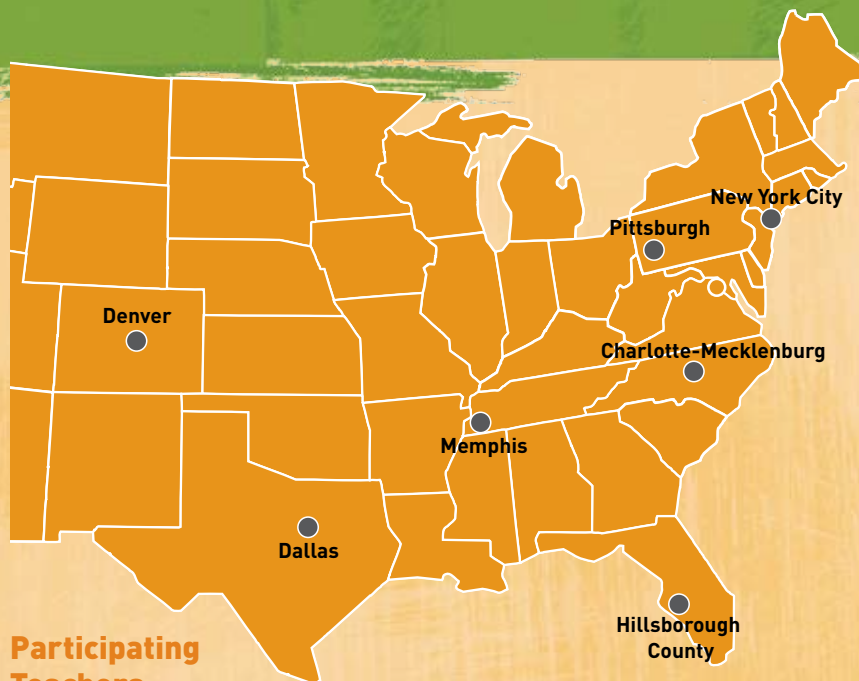
Participating Teachers

Charlotte-Mecklenberg	465	Memphis	444
Dallas	315	New York City	728
Denver	130	Pittsburgh	196
Hillsborough County	741	TOTAL	3,019

English language arts (ELA) in grades 4–8, Algebra I, grade 9 English, and high school biology.

All MET project teachers have agreed to have the following data collected and analyzed:

- students' performance on standardized state and supplemental assessments

- video-based classroom observation (four lessons per teacher per year) and teachers' reflections on these lessons

- teachers' pedagogical content knowledge—an assessment of a teacher's ability to recognize and diagnose students' misunderstandings of the lessons

- students' perceptions of the instructional environment in the classroom

- teachers' perceptions of the working conditions and instructional support at their schools

None of the individual teacher-level data collected as part of this project will be shared with principals or other school or district personnel. If it is determined that aggregated data would help school districts, and if such data can be provided without identifying individual teachers, then the data will be provided at the districts' request.

Measure 1: Student achievement gains on assessments

Student achievement is being measured in two ways—through existing state assessments, designed to assess student progress on the state curriculum for accountability purposes, and supplemental assessments, designed to assess higher-order conceptual understanding. Together, these two forms of assessment mitigate the widespread concern that evaluation systems primarily measure test-taking skills rather than higher-order thinking and therefore encourage "teaching to the test." The supplemental assessments are Stanford 9 Open-Ended Reading Assessment in grades 4 through 8, Balanced Assessment in Mathematics (BAM) in grades 4 through 8, and the ACT QualityCore series for Algebra I, English 9, and Biology.

Measure 2: Classroom observations and teacher reflections

One of the most difficult challenges in designing the MET project was to find a way to observe more than 20,000 lessons at a reasonable cost. Videotaping seemed like a reasonable option, but for videotaped lessons to become a viable approach for observing classrooms, the

project had to overcome several technical challenges. The solution, engineered by Teachscape, involves panoramic digital video cameras that require minimal training to set up, are operated remotely by the individual teachers, and do not require a cameraperson. After class, participating teachers upload video lessons to a secure Internet site.

The participating teachers offer commentary on their lessons (e.g., specifying the learning objective). Then, trained raters score the lesson based on classroom observation protocols developed by leading academics and professional development experts. The raters examine everything from the teacher's ability to establish a positive learning climate and manage his/her classroom to the ability to explain concepts and provide useful feedback to students. The Educational Testing Service (ETS) manages the lesson-scoring process. Personnel from ETS have trained raters to accurately score lessons using the following five observation protocols:

- Classroom Assessment Scoring System (CLASS), developed by Robert Pianta, University of Virginia

- Framework for Teaching, developed by Charlotte Danielson

- Mathematical Quality of Instruction (MQI), developed by Heather Hill, Harvard University, and Deborah Loewenberg Ball, University of Michigan

- Protocol for Language Arts Teaching Observations (PLATO), developed by Pam Grossman, Stanford University

- Quality Science Teaching (QST) Instrument, developed by Raymond Pecheone, Stanford University

A subset of the videos also are being scored using an observational protocol developed by the National Board for Professional Teaching Standards (NBPTS).

Teacher Advisory Panel

The MET project is guided by our Teacher Advisory Panel (TAP), a group of 21 classroom teachers who advise on the research tools, implementation strategies and challenges, and emerging findings. This diverse group of teachers represents all geographic regions, grade levels, and subject areas. The teachers have from 3 to 20-plus years' experience in the classroom.

When they first met, the TAP members spoke openly about the perfunctory nature of evaluation and their hunger for feedback to help them develop their craft. One teacher mentioned that her colleagues "welcomed their fifth year evaluation because it keeps them from having to go through the evaluation process again for five more years." Another teacher questioned whether the 40-year-old evaluation instrument in use in his district was attuned to the needs of this generation's students and teachers. Two of the advisory panel teachers shared their positive experience of receiving feedback that helped them become more expert.

The advisory panel showed great enthusiasm for the proposed MET project processes and tools, especially the video capture and viewing system and the students' assessment of the instructional environment. One TAP member said, "Some kids will use it as a gotcha, but most will be honest. I really want to know what they think." Some advisory panel members previously had used video for feedback purposes and were quick to point out the initial discomfort ("Who likes seeing themselves on video?") but recognized the value of the questions it raised about their own practice: "I couldn't believe I said what I said in that way. It gave me lots to think about."

Measure 3: Teachers' pedagogical content knowledge

ETS, in collaboration with researchers at the University of Michigan's Learning Mathematics for Teaching Project, has developed an assessment to measure teachers' general, specialized, and pedagogical content knowledge. Expert teachers can identify errors in student reasoning and use this knowledge to develop a strategy to correct the errors and strengthen student understanding. These assessments focus on specialized knowledge that teachers use to interpret student responses, choose instructional strategies, detect and address student errors, select models to illustrate particular instructional objectives, and understand the special instructional challenges faced by English language learners.

Measure 4: Student perceptions of the classroom instructional environment

Students also report their experiences of the classroom instructional environment. The Tripod survey instrument, developed by Harvard researcher Ron Ferguson and administered by Cambridge Education, will assess the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth. The survey will ask students in the classrooms of the more than 3,000 participating teachers if they agree or disagree with a variety of statements, including: "My teacher knows when the class understands, and when we do not"; "My teacher has several good ways to explain each topic that we cover in this class"; and "When I turn in my work, my teacher gives me useful feedback that helps me improve." Such questions offer students the chance to give feedback on specific aspects of a teacher's practice so that teachers can ultimately improve their use of class time, the quality of the comments they give on homework, their pedagogical practices, and their relationships with their students.

Measure 5: Teachers' perceptions of working conditions and instructional support at their schools

Teachers also complete a survey, developed by the New Teacher Center, about working conditions, school environment, and the instructional support they receive in their schools. Indicators include whether teachers are encouraged to try new approaches to improve instruction or whether they receive an appropriate amount of professional development. The survey is intended to give teachers a voice in providing feedback on the quality of instructional support they receive. The results potentially could be incorporated into measuring the effectiveness of principals in supporting effective instruction.

Building and Validating a Composite Model

The MET project is analyzing its data in three stages. The first of these stages already has begun. We are using three years of historical data on student performance, student demographics, and teacher characteristics (such as degrees, certification, licensing scores, tenure, district performance review ratings, years of experience, and NBPTS status) to estimate each participating teacher's impact on student achievement gains. These data will serve as a benchmark and help determine the extent to which a teacher's impact on student performance in 2009–10 compares to past years.

In the second stage, researchers from RAND will combine data from each of the MET project measures to form a composite indicator of effective teaching. We will assign a weight to each measure (classroom observations, teacher knowledge, student perceptions, and teacher perceptions) based on the result of analyses that indicate how much each weight contributes to predicting student learning gains.

In the third stage, we will test whether those teachers whose performance was rated most highly during year 1 (2009–10) actually produce larger student achievement gains than their colleagues in year 2 (2010–11), or whether those teachers simply appear to be more effective than their colleagues because of the composition of their classes or other factors.

We will be testing whether students who have teachers with the highest composite scores actually show the most improvement. Analysis of year 2 data also will surface specific measures (e.g., student perceptions) that are more predictive than others and should therefore have their weight adjusted. Conversely, the year 2 data may illuminate factors that turn out not to be particularly predictive of student success. In either case, the research will

provide unique insights. Teaching is, after all, multidimensional, so we hope that the composite measure will be a stable predictor of student achievement gains in a particular teacher's classroom.

Although the analysis of achievement gains during year 1 will include statistical controls for prior achievement, students may differ in other hard-to-measure ways (such as behavior or engagement in learning). The only way to control for all the ways in which students differ is through random assignment, so teachers participating in the MET project have signed up as groups of three or more colleagues working in the same school, same grade, and same subjects. Once schools have drawn up rosters of students in their grades and subjects, teachers in each group have agreed to the random

assignment of the students they will teach in year 2 of the study. Researchers will study differences in student achievement gains within each of those groupings to see if the students assigned to the teachers identified as "more effective" actually outperform the students assigned to the "less effective" teachers. Random assignment means that there should be no differences— measured or unmeasured—in the baseline characteristics of the students assigned to the "more effective" or "less effective" teachers.

Sharing the Research

The bottom line: Better student achievement will require better teaching. The MET project will help pinpoint what that looks like in practice … and then broadly share our findings and recommendations with practitioners and policymakers across the country.

With Participating Teachers and Districts

Our immediate priority is to provide regular input and feedback to project participants. We are conducting weekly webinars on data collection and implementation with each district's project manager, who then disseminates relevant information to all participating teachers and principals. Participating teachers also will see their own classroom videos and can get access to their schools' working condition survey results if the school's response rate is greater than 50 percent.

Although we will not provide districts with data about individual teachers, we will convene district partners in summer 2010 to share preliminary findings and tools and facilitate a conversation about how they want to use these results and tools to improve teacher support and evaluation in their schools. In addition, we likely will host a final district convening at the end of the project.

With Other Practitioners and Policymakers

As the MET project progresses and we learn more about teacher effectiveness, how to measure it and how to increase the quality of information and tools used within state, district, and school evaluation systems, we will share a series of publications and tools more broadly. Our reports will address the following topics:

- interim findings and results

- study design, methods, and empirical analyses

- teacher observational protocols, training, and scoring requirements

- final findings and results (fall 2011)

- implementation guides and data requirements, showing how to use the composite measure and gather and store the data

"This national research study is going to help all of us in public education learn about great teaching because it's going to study real teachers in real classrooms."

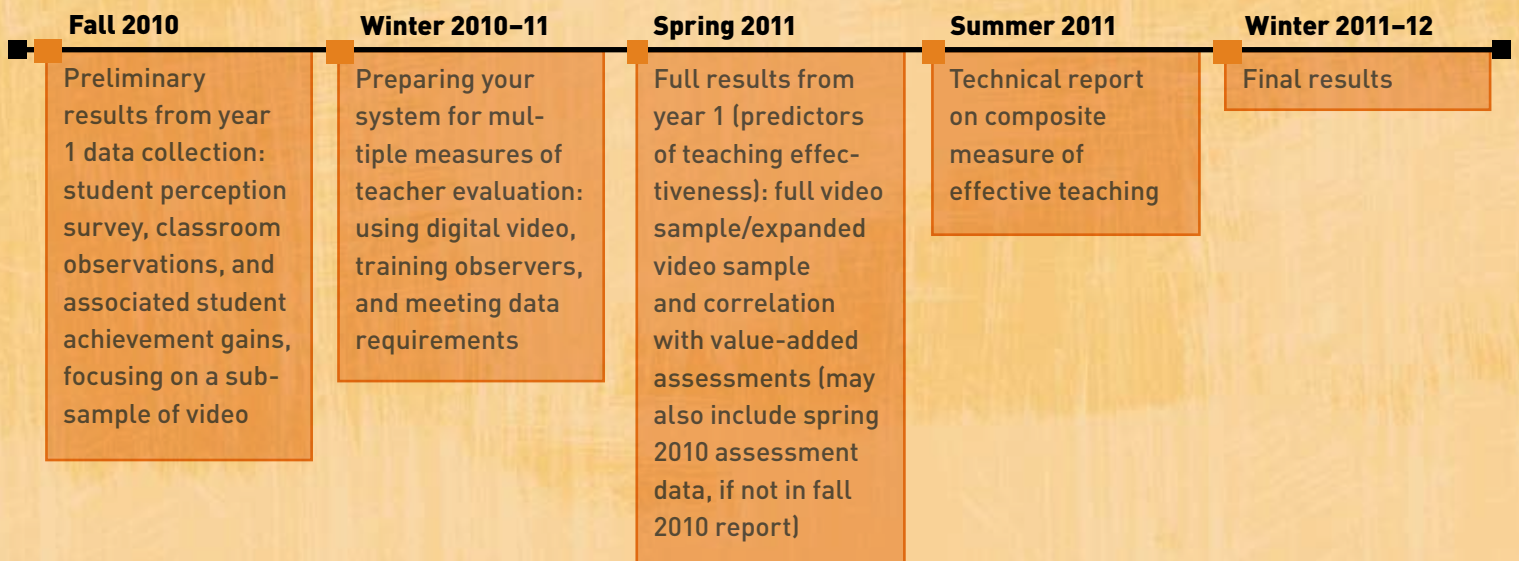**—Pete Gorman,** *Superintendent, Charlotte-Mecklenburg (NC) Schools*

In addition, we will publish a toolkit for measuring effective teaching, which will include:

- a student survey instrument

- a teacher survey instrument

- advice and a process for training raters to make consistent observations of classroom practice

- advice on how to set up low-cost, good-quality video-capture devices, storage capacity, and retrieval software

Our method of videotaped teacher observation holds great promise for both teacher evaluation and professional development. The use of digital video makes it possible to have multiple professionals look at the same evidence, thereby making ratings less subjective. Moreover, teachers will use the videos for self-reflection, feedback from peers, and tracking professional growth. Finally, the existence of video makes it much easier to share the work of exemplary teachers.

By determining exactly what measures predict the biggest student achievement gains, the MET project will give teachers the feedback (including exemplary practices) they need to improve. In addition, a greater understanding about which teaching practices, skills, and knowledge positively impact student learning will allow states and districts to develop teacher evaluation systems that will help strengthen all aspects of teaching—from recruitment through retention.

MET Project Implementation Timeline

Fall 2010	Winter 2010–11	Spring 2011	Summer 2011	Winter 2011–12
Preliminary results from year 1 data collection: student perception survey, classroom observations, and associated student achievement gains, focusing on a sub-sample of video	Preparing your system for multiple measures of teacher evaluation: using digital video, training observers, and meeting data requirements	Full results from year 1 (predictors of teaching effectiveness): full video sample/expanded video sample and correlation with value-added assessments (may also include spring 2010 assessment data, if not in fall 2010 report)	Technical report on composite measure of effective teaching	Final results

Endnotes

1 Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica*, Vol. 73, No. 2 (March 2005), pages 417–458. http://edpro.stanford.edu/Hanushek/admin/pages/files/uploads/teachers.econometrica.pdf

2 Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, "Identifying Effective Teachers Using Performance on the Job," Hamilton Project Discussion Paper (March 2006), The Brookings Institution. www.brookings.edu/views/Papers/200604hamilton_1.pdf

3 Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," Working Paper 11 (October 2007), National Center for the Analysis of Longitudinal Data in Education Research. www.caldercenter.org/PDF/1001104_Teacher_Credentials_HighSchool.pdf

4 "Highlights from TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context" (September 2009), National Center for Education Statistics, Institute of Education Sciences. http://nces.ed.gov/pubs2009/2009001.pdf

5 D. Weisberg et al. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (2009), The New Teacher Project. http://widgeteffect.org/downloads/TheWidgetEffect.pdf 2009

6 Donald J. Boyd et al. (2006), "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy,* Vol. 1, No. 2 (Spring 2006), pages 176–216. www.mitpressjournals.org/doi/pdf/10.1162/edfp.2006.1.2.176

7 States and districts will choose the outcomes most relevant for their systems. If new tests or different ways of assessing student learning become available (i.e., portfolios or other tools), the indicators of teaching effectiveness would need to be validated against these new assessment methods.

BILL & MELINDA
GATES *foundation*