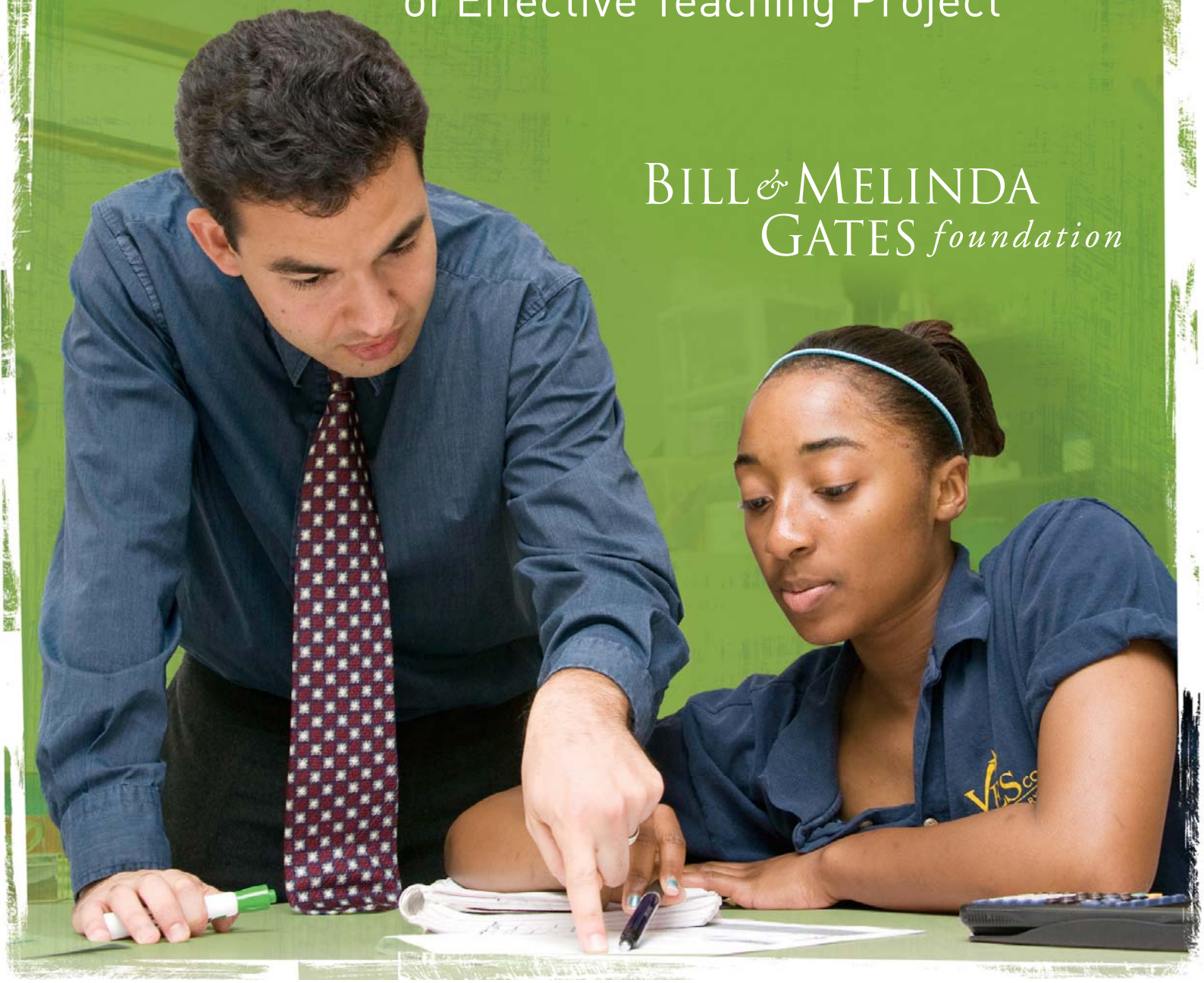
MET project

RESEARCH PAPER

Learning about Teaching

Initial Findings from the Measures of Effective Teaching Project



BILL & MELINDA GATES *foundation*

About the Measures of Effective Teaching Project

In fall 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to test new approaches to measuring effective teaching. 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 measuring teacher effectiveness that can be used for a variety of purposes, including feedback, development, and continuous improvement. The project includes nearly 3000 teachers who volunteered to help us identify a better approach to teacher development and evaluation, located in six predominantly urban school districts across the country: Charlotte-Mecklenburg Schools, Dallas Independent School District, Denver Public Schools, Hillsborough County Public Schools (including Tampa, Florida), Memphis City Schools, and the New York City Department of Education. As part of the project, multiple data sources are being collected and analyzed over two school years, including student achievement gains on state assessments and supplemental assessments designed to assess higher-order conceptual understanding; classroom observations and teacher reflections on their practice; assessments of teachers' pedagogical content knowledge; student perceptions of the classroom instructional environment; and teachers' perceptions of working conditions and instructional support at their schools.

The project is directed by Thomas J. Kane, Deputy Director and Steven Cantrell, Senior Program Officer at the Bill & Melinda Gates Foundation. Our lead research partners include:

- Mark Atkinson, Teachscape
- Nancy Caldwell, Westat
- Charlotte Danielson, The Danielson Group
- Ron Ferguson, Harvard University
- Drew Gitomer, Educational Testing Service
- Pam Grossman, Stanford University
- Heather Hill, Harvard University
- Eric Hirsch, New Teacher Center
- Dan McCaffrey, RAND
- Catherine McClellan, Educational Testing Service
- Roy Pea, Stanford University
- Raymond Pecheone, Stanford University
- Geoffrey Phelps, Educational Testing Service
- Robert Pianta, University of Virginia
- Rob Ramsdell, Cambridge Education
- Doug Staiger, Dartmouth College
- John Winn, National Math and Science Initiative

Introduction

For four decades, educational researchers have confirmed what many parents know: children's academic progress depends heavily on the talent and skills of the teacher leading their classroom. Although parents may fret over their choice of school, research suggests that their child's teacher assignment in that school matters a lot more.

And yet, in most public school districts, individual teachers receive little feedback on the work they do. Almost everywhere, teacher evaluation is a perfunctory exercise. In too many schools principals go through the motions of visiting classrooms, checklist in hand. In the end, virtually all teachers receive the same "satisfactory" rating.[1]

The costs of this neglect are enormous. Novice teachers' skills plateau far too early without the feedback they need to grow. 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 we ask of them—granting tenure to beginning teachers still 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 they do in the classroom? If teachers are producing dramatically different results, why don't we provide them with that feedback and trust them to respond to it?

Resolving the contradiction will require new tools for gaining insight into teachers' practice, new ways to diagnose their strengths and weaknesses and new approaches to developing teachers. In the Fall of 2009, the Bill & Melinda Gates Foundation launched the Measures of Effective Teaching (MET) project to test new approaches to identifying effective teaching. The goal of the project is to improve the quality of information about teaching effectiveness, to help build fair and reliable systems for teacher observation and feedback.

OUR PARTNERS

Although funded by the Bill & Melinda Gates Foundation, the MET project is led by more than a dozen organizations, including 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, the National Math and Science Initiative, and the New Teacher Center), and other educational consultants (Cambridge Education, Teachscape, Westat, and the Danielson Group).

In addition, the National Board for Professional Teaching Standards and Teach for America have encouraged their members to participate. The American Federation of Teachers and the National Education Association have been engaged in the project. Indeed, their local leaders actively helped recruit teachers.

1 The 2009 New Teacher Project study, *The Widget Effect*, found that evaluation systems with two ratings, "satisfactory" and "unsatisfactory," 99 percent of teachers earned a satisfactory. In evaluation systems with more than two ratings, 94 percent of teachers received one of the top two ratings and less than one percent were rated unsatisfactory.

Yet, our most vital partners are the nearly 3000 teacher volunteers in six school districts around the country who volunteered for the project. They did so because of their commitment to the profession and their desire to develop better tools for feedback and growth.

The six districts hosting the project are all predominantly urban districts, spread across the country: Charlotte-Mecklenburg Schools, Dallas Independent School District, Denver Public Schools, Hillsborough County Public Schools (including Tampa, Florida), Memphis City Schools, and the New York City Department of Education.

THE THREE PREMISES OF THE MET PROJECT

The MET project is based on three simple premises:

First, whenever feasible, a teacher's evaluation should include his or her students' achievement gains.

Some raise legitimate concerns about whether student achievement gains measure all of what we seek from teaching. Of course, they're right. Every parent wants their children to build social skills and to acquire a love of learning. Likewise, our diverse society needs children who are tolerant. However, these goals are not necessarily at odds with achievement on state tests. For instance, it may be that an effective teacher succeeds by inspiring a love of learning, or by coaching children to work together effectively. We will be testing these hypotheses in future reports, using the data from our student surveys. For example, it may be possible to add measures of student engagement as additional outcome measures. This would be particularly useful in grades and subjects where testing is not feasible.Others have raised separate concerns about whether "value-added" estimates (which use statistical methods to identify the impact of teachers and schools by adjusting for students' prior achievement and other measured characteristics) are "biased" (Rothstein, 2010). They point out that some teachers may be assigned students that are systematically different in other ways—such as motivation or parental engagement—which affect their ultimate performance but are not adequately captured by prior achievement measures. As we describe below, our study aspires to resolve that question with a report next winter. At that time, we will be testing whether value-added measures accurately predict student achievement following random assignment of teachers to classrooms (within a school, grade and subject). However, in the interim, there is little evidence to suggest that value-added measures are so biased as to be directionally misleading. On the contrary, in a small sample of teachers assigned to specific rosters by lottery, Kane and Staiger (2008) could not reject that there was no bias and that the value-added measures approximated "causal" teacher effects on student achievement. Moreover, a recent re-analysis of an experiment designed to test classroom size, but which also randomly assigned students to teachers, reported teacher effects on student achievement which were, in fact, larger than many of those reported in value-added analyses (Nye, Konstantopoulos and Hedges, 2004). Value-added measures do seem to convey information about a teacher's impact. However, evidence of bias at the end of this year may require scaling down (or up) the value-added measures themselves. But that's largely a matter of determining how much weight should be attached to value-added as one of multiple measures of teacher effectiveness.

Second, any additional components of the evaluation (e.g., classroom observations, student feedback) should be demonstrably related to student achievement gains.

The second principle is fundamental, especially given that most teachers are receiving the same "satisfactory" rating now. If school districts and states simply give principals a new checklist to fill out during their classroom visits little will change. The only way to be confident that the new feedback is pointing teachers in the right direction—toward improved student achievement—is to regularly confirm that those teachers who receive higher ratings actually achieve greater student achievement gains on average. Even a great system can be implemented poorly or gradually succumb to "grade inflation". Benchmarking against student achievement gains is the best way to know when the evaluation system is getting closer to the truth—or regressing. Accordingly, in our own work, we will be testing whether student perceptions, classroom observations and assessments of teachers' pedagogical content knowledge are aligned with value-added measures.

Third, the measure should include feedback on specific aspects of a teacher's practice to support teacher growth and development.

Any measure of teacher effectiveness should support the continued growth of teachers, by providing actionable data on specific strengths and weaknesses. Even if value-added measures are valid measures of a teacher's impact on student learning, they provide little guidance to teachers (or their supervisors) on what they need to do to improve. Therefore, our goal is to identify a package of measures, including student feedback and classroom observations, which would not only help identify effective teaching, but also point all teachers to the areas where they need to become more effective teachers themselves.

The Measures

To limit the need for extensive additional testing, the MET project started with grades and subjects where most states currently test students. We included those teaching mathematics or English language arts in grades 4 through 8. In addition, we added three courses which serve as gateways for high school students, where some states are using end-of-course tests: Algebra I, grade 9 English, and biology.

The following data are being collected in their classrooms.

Measure 1: Student achievement gains on different assessments

Student achievement is being measured in two ways, with existing state assessments and with three supplemental assessments. The latter are designed to assess higher-order conceptual understanding. By combining the state tests and the supplemental tests, we plan to test whether the teachers who are successful in supporting student gains on the state tests are also seeing gains on the supplemental assessments. 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 was an intriguing alternative to in-person observations (especially given our aspiration to test multiple rubrics), but the project had to overcome several technical challenges: tracking both students and a non-stationary teacher without having another adult in the classroom pointing the camera and distracting children, sufficient resolution to read a teacher's writing on a board or projector screen, and sufficient audio quality to hear teachers and students. 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.[2] After class, participating teachers upload video lessons to a secure Internet site, where they are able to view themselves teaching (often for the first time).

In addition, the participating teachers offer limited commentary on their lessons (e.g., specifying the learning objective). Trained raters are scoring the lessons 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) is managing 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 Bob Pianta and Bridget Hamre, University of Virginia

2 Similar cameras have been developed by other suppliers, such as thereNow (www.therenow.net). A commercial version of the camera used in the MET project is available from Kogeto. (www.kogeto.com).

- Framework for Teaching, developed by Charlotte Danielson (2007)
- 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 is also being scored by the National Board for Professional Teaching Standards (NBPTS). In addition, the National Math and Science Initiative (NMSI) is scoring a subset of videos using the UTeach Observation Protocol (UTOP) for evaluating math instruction, developed and field tested over three years by the UTeach program at the University of Texas at Austin.

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' knowledge for teaching—not just their content knowledge. Expert teachers should be able to identify common errors in student reasoning and use this knowledge to develop a strategy to correct the errors and strengthen student understanding. The new assessments to be administered this year 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 in the MET classrooms were asked to report their perceptions of the classroom instructional environment. The Tripod survey instrument, developed by Harvard researcher Ron Ferguson and administered by Cambridge Education, assesses the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth. The survey asks students in the each of the MET classrooms 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."

The goal is not to conduct a popularity contest for teachers. Rather, students are asked to give feedback on *specific aspects* of a teacher's practice, so that teachers can improve their use of class time, the quality of the comments they give on homework, their pedagogical practices, or 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. Although we have not yet had a chance to analyze those data for the current report, they will be included in future analyses.

Stages of Analysis

The MET project will be issuing four reports, starting with this one. In this preliminary report of findings from the first year, we focus on mathematics and English language arts teachers, in grades 4 through 8, in five of the six districts. (The student scores on the state tests were not available in time to include teachers in Memphis). We report the relationships across a variety of measures of effective teaching, using data from one group of students or school year to identify teachers likely to witness success with another group of students or during another school year.

At this point, we have classroom observation scores for a small subset (less than 10 percent) of the lessons collected last year. Given the importance of those findings, we will issue a more complete report in the spring of 2011, including a much larger sample of videos. Our aim is to test various approaches to classroom observations.

Third, late in the summer of 2011, researchers from RAND will combine data from each of the MET project measures to form a "composite indicator" of effective teaching. That report will assign a weight to each measure (classroom observations, teacher knowledge, and student perceptions) based on the result of analyses indicating how helpful each is in identifying teachers likely to produce exemplary student learning gains.

Our goal is to identify effective teachers and effective teaching practices. To do so, we need to isolate the results of effective teaching from the fruits of a favorable classroom composition. It may well be easier to use certain teaching practices or to garner enthusiastic responses from students if one's students show up in class eager to learn. If that's the case, we would be in danger of confusing the effects of teachers with the effects of classroom characteristics.

Like virtually all other research on the topic of effective teaching, we use statistical controls to account for differences in students' entering characteristics. But it is always possible to identify variables for which one has not controlled. The only way to resolve the question of the degree of bias in our current measures is through random assignment. As a result, teachers participating in the MET project signed up in groups of two or more colleagues working in the same school, same grade, and same subjects. During the spring and summer of 2010, schools drew up a set of rosters of students in each of those grades and subjects and submitted them to our partners at RAND. RAND then randomly assigned classroom rosters within the groups of teachers in a given grade and subject (so that no teacher was asked to teach in a grade, subject or school where they did not teach during year one). Within each group of teachers in a school, grade and subject, teachers effectively drew straws to determine which group of students they would teach this year.

At the end of the current school year, we will study differences in student achievement gains within each of those groupings to see if the students assigned to the teachers identified using year one data as "more effective" actually outperform the students assigned to the "less effective" teachers. We will look at differences in student achievement gains within each of those groups and then aggregate up those differences for "more effective" and "less effective" teachers. Following random assignment, there should be no differences—measured or unmeasured—in the prior characteristics of the students assigned to "more effective" or "less effective" teachers as a group. If the students assigned to teachers who were identified as "more effective" outperform those assigned to "less effective" teachers, we can resolve any lingering doubts about whether the achievement differences represent the effect of teachers or unmeasured characteristics of their classes.

Better student achievement will require better teaching. The MET project is testing novel ways to recognize effective teaching. We hope the results will be used to provide better feedback to teachers and establish better ways to help teachers develop.

What We're Learning So Far

Before describing the measures and analysis in more detail, we briefly summarize our findings so far.

- In every grade and subject, a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years. A teacher's value-added fluctuates from year-to-year and from class-to-class, as succeeding cohorts of students move through their classrooms. However, that volatility is not so large as to undercut the usefulness of value-added as an indicator (imperfect, but still informative) of future performance.

The teachers who lead students to achievement gains in one year or in one class tend to do so in other years and other classes.

- Teachers with high value-added on state tests tend to promote deeper conceptual understanding as well.

Many are concerned that high value-added teachers are simply coaching children to do well on state tests. In the long run, it would do students little good to score well on state tests if they fail to understand key concepts. However, in our analysis so far, that does not seem to be the case. Indeed, the teachers who are producing gains on the state tests are generally also promoting deeper conceptual understanding among their students. In mathematics, for instance, after adjusting for measurement error, the correlation between teacher effects on the state math test and on the Balanced Assessment in Mathematics was moderately large, .54.

- Teachers have larger effects on math achievement than on achievement in reading or English Language Arts, at least as measured on state assessments.

Many researchers have reported a similar result: teachers seem to have a larger influence on math performance than English Language Arts performance. A common interpretation is that families have more profound effects on children's reading and verbal performance than teachers. However, the finding may also be due to limitations of the current state ELA tests (which typically consist of multiple-choice questions of reading comprehension). When using the Stanford 9 Open-Ended assessment (which requires youth to provide written responses), we find teacher effects comparable to those found in mathematics. We will be studying this question further in the coming months, by studying teacher effects on different types of test items. However, if future work confirms our initial findings with the open-ended assessment, it would imply that the new literacy assessments, which are being designed to assess the new common core standards, may be more sensitive to instructional effects than current state ELA tests.

- Student perceptions of a given teacher's strengths and weaknesses are consistent across the different groups of students they teach. Moreover, students seem to know effective teaching when they experience it: student perceptions in one class are related to the achievement gains in other classes taught by the same teacher. Most important are students' perception of a teacher's ability to control a classroom and to challenge students with rigorous work.

While student feedback is widely used in higher education, it is rare for elementary and secondary schools to ask youth about their experiences in the classroom. Nevertheless, soliciting student feedback is potentially attractive for a number of reasons: the questions themselves enjoy immediate legitimacy with teachers, school leaders and parents; it is an inexpensive way to supplement other more costly indicators, such as classroom observations; and the questionnaires can be extended to non-tested grades and subjects quickly. Our preliminary results suggest that the student questionnaires would be a valuable complement to other performance measures.

Classroom observations are the most common form of evaluation today. As a result, our goal is to test several different approaches to identifying effective teaching practices in the classroom. In our work so far, we have some promising findings suggesting that classroom observations are positively related to student achievement gains. However, because less than 10 percent of the videos have been scored, we will be waiting until April to release results on the classroom observation methods.

MEASURING TEACHER-LEVEL VALUE-ADDED

In order to put the measures of student achievement on a similar footing, we first standardized test scores to have a mean of 0 and a standard deviation of 1 (for each district, subject year and grade level). We then estimated a statistical model controlling for each student's test score in that subject from the prior year, a set of student characteristics and the mean prior test score and the mean student characteristics in the specific course section or class which the student attends. (We provide more details in the Technical Appendix.) The student characteristics varied somewhat by district (depending upon what was available), but typically included student demographics, free or reduced price lunch, ELL status and special education status[3]. The statistical model produces an "expected" achievement for each student based on his or her starting point and the starting point of his or her peers in class. Some students "underperformed" relative to that expectation and some students "overperformed". In our analysis, *a teacher's "value-added" is the mean difference, across all tested students in a classroom with a prior year achievement test score, between their actual and expected performance at the end of the year.* If the average student in the classroom outperformed students elsewhere who had similar performance on last year's test, similar demographic and program participation codes—and classmates with similar prior year test scores and other characteristics—we infer a positive value-added, or positive achievement gain, attributable to the teacher.

Using this method, we generated value-added estimates on the state assessments and the supplemental assessments for up to two course sections or classrooms teachers taught during 2009-10. We also calculated value-added estimates for teachers on state math and ELA test scores using similar data we obtained from the districts from the 2008-09 school year. (To be part of the MET project, a district was required to have some historical data linking students and teachers.)

3 The student-level covariates used in the regressions included, in Charlotte-Mecklenburg: race, ELL status, age, gender, special education, gifted status; in Dallas: race, ELL, age, gender, special education, free or reduced lunch; in Denver: race, age and gender; in Hillsborough: race, ELL, age, special education, gifted status, and free or reduced lunch; in NYC: race, ELL, gender, special education, free or reduced lunch. Differences in covariates across districts may reduce the reliability of the value added estimates.

In addition to state tests, students in participating classes took a supplemental performance assessment in spring 2010. Students in grades 4-8 math classes took the Balanced Assessment in Mathematics, while students in grades 4-8 English language arts classes took the SAT 9 Open-Ended Reading assessment. We chose these two tests because they included cognitively demanding content, they were reasonably well-aligned with the curriculum in the six states, had high levels of reliability, and had evidence of fairness to members of different groups of students.

Balanced Assessment in Mathematics (BAM): Each of the test forms for the Balanced Assessment in Mathematics (BAM) includes four to five tasks and requires 50-60 minutes to complete. Because of the small number of tasks on each test form, however, we were concerned about the content coverage in each teacher's classroom. As a result, we used three different forms of the BAM—from the relevant grade levels in 2003, 2004 and 2005—in each classroom. In comparison to many other assessments, BAM is considered to be more cognitively demanding and measures higher order reasoning skills using question formats that are quite different from those in most state mathematics achievement tests. There is also some evidence that BAM is more instructionally sensitive to the effects of reform-oriented instruction than a more traditional test (ITBS). Appendix 1 includes some sample items from the BAM assessment.

SAT 9 Reading Open-Ended Test: The Stanford 9 Open-Ended (OE) Reading assessment contains nine open-ended tasks and takes 50 minutes to complete. The primary difference between the Stanford 9 OE and traditional state reading assessments is the exclusive use of open-ended items tied to extended reading passages. Each form of the assessment consists of a narrative reading selection followed by nine questions. Students are required to not only answer the questions but also to explain their answers. Sample items from the Stanford 9 OE exam are available in Appendix 2.

MEASURING STUDENT PERCEPTIONS

College administrators rarely evaluate teaching by sitting in classrooms—as is the norm in K–12 schools. Rather, they rely on confidential student evaluations. Organizers of the MET project wondered whether such information could be helpful in elementary and secondary schools, to supplement other forms of feedback.

The MET student perceptions survey is based on a decade of work by the Tripod Project for School Improvement. Tripod was founded by Ronald F. Ferguson of Harvard University and refined in consultation with K-12 teachers and administrators in Shaker Heights, Ohio, and member districts of the Minority Student Achievement Network. For the MET project, the Tripod surveys are conducted either online or on paper, at the choice of the participating school. For online surveys, each student is given a ticket with a unique identification code to access the web site. For the paper version, each form is pre-coded with a bar code identifier. When a student completes a paper survey, he or she seals it in a thick, non-transparent envelope. The envelope is opened only at a location where workers scan the forms to capture the data. These precautions are intended to ensure that students feel comfortable providing their honest feedback, without the fear that their teacher will tie the feedback to them.

The Tripod questions are gathered under seven headings, or constructs, called the Seven C's. The seven are: Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate. Each of the C's is measured using multiple survey items. Tables 1 and 2 provides a list of the items used to measure each of the Seven C's on the elementary and secondary survey respectively. The indices for the Seven C's have proven highly reliable—in

Table 1. Rates of Agreement at the Classroom Level to Tripod Survey Items: Elementary

CARE	25TH PERCENTILE	75TH PERCENTILE
I like the way my teacher treats me when I need help.	76	94
My teacher is nice to me when I ask questions.	74	93
My teacher in this class makes me feel that he/she really cares about me.	70	91
If I am sad or angry, my teacher helps me feel better.	50	76
The teacher in this class encourages me to do my best.	82	95
My teacher seems to know if something is bothering me.	50	70
My teacher gives us time to explain our ideas.	68	88
CONTROL		
My classmates behave the way my teacher wants them to.	23	53
Our class stays busy and does not waste time.	44	71
Students behave so badly in this class that it slows down our learning.	12	41
Everybody knows what they should be doing and learning in this class.	69	87
CLARIFY		
My teacher explains things in very orderly ways.	67	85
In this class, we learn to correct our mistakes.	83	95
My teacher explains difficult things clearly.	75	90
My teacher has several good ways to explain each topic that we cover in this class.	72	89
I understand what I am supposed to be learning in this class.	76	91
My teacher knows when the class understands, and when we do not.	67	85
This class is neat—everything has a place and things are easy to find.	55	80
If you don't understand something, my teacher explains it another way.	75	91
CHALLENGE		
My teacher pushes us to think hard about things we read.	57	81
My teacher pushes everybody to work hard.	67	87
In this class we have to think hard about the writing we do.	65	85
In this class, my teacher accepts nothing less than our full effort.	75	90
CAPTIVATE		
School work is interesting.	40	67
We have interesting homework.	33	58
Homework helps me learn.	65	86
School work is not very enjoyable. (Do you agree?)	19	35
CONFER		
When he/she is teaching us, my teacher asks us whether we understand.	76	92
My teacher asks questions to be sure we are following along when he/she is teaching.	81	94
My teacher checks to make sure we understand what he/she is teaching us.	80	94
My teacher tells us what we are learning and why.	71	89
My teacher wants us to share our thoughts.	53	78
Students speak up and share their ideas about class work.	47	71
My teacher wants me to explain my answers—why I think what I think.	70	88
CONSOLIDATE		
My teacher takes the time to summarize what we learn each day.	48	73
When my teacher marks my work, he/she writes on my papers to help me understand.	52	75

Note: For each question, a quarter of classrooms had a lower percentage of students agreeing than the 25th percentile and another quarter of classrooms had rates of agreement higher than the 75th percentile. There were 963 elementary classrooms with more than 5 students responding.

Table 2. Rates of Agreement at the Classroom Level to Tripod Survey Items: Secondary

CARE	25TH PERCENTILE	75TH PERCENTILE
My teacher in this class makes me feel that s/he really cares about me.	40	73
My teacher seems to know if something is bothering me.	22	50
My teacher really tries to understand how students feel about things.	35	68
CONTROL		
Student behavior in this class is under control.	30	67
I hate the way that students behave in this class.	10	32
Student behavior in this class makes the teacher angry.	17	50
Student behavior in this class is a problem.	9	37
My classmates behave the way my teacher wants them to.	20	57
Students in this class treat the teacher with respect.	33	79
Our class stays busy and doesn't waste time.	36	69
CLARIFY		
If you don't understand something, my teacher explains it another way.	60	86
My teacher knows when the class understands, and when we do not.	50	77
When s/he is teaching us, my teacher thinks we understand even when we don't.	9	27
My teacher has several good ways to explain each topic that we cover in this class.	53	82
My teacher explains difficult things clearly.	50	79
CHALLENGE		
My teacher asks questions to be sure we are following along when s/he is teaching.	75	93
My teacher asks students to explain more about answers they give.	63	86
In this class, my teacher accepts nothing less than our full effort.	53	81
My teacher doesn't let people give up when the work gets hard.	56	83
My teacher wants us to use our thinking skills, not just memorize things.	63	85
My teacher wants me to explain my answers—why I think what I think.	59	83
In this class, we learn a lot almost every day.	52	81
In this class, we learn to correct our mistakes.	56	83
CAPTIVATE		
This class does not keep my attention—I get bored.	14	36
My teacher makes learning enjoyable.	33	72
My teacher makes lessons interesting.	33	70
I like the ways we learn in this class.	47	81
CONFER		
My teacher wants us to share our thoughts.	47	79
Students get to decide how activities are done in this class.	5	20
My teacher gives us time to explain our ideas.	43	73
Students speak up and share their ideas about class work.	40	68
My teacher respects my ideas and suggestions.	46	75
CONSOLIDATE		
My teacher takes the time to summarize what we learn each day.	38	67
My teacher checks to make sure we understand what s/he is teaching us.	58	86
We get helpful comments to let us know what we did wrong on assignments.	45	74
The comments that I get on my work in this class help me understand how to improve.	46	74

Note: For each question, a quarter of classrooms had a lower percentage of students agreeing than the 25th percentile and another quarter of classrooms had rates of agreement higher than the 75th percentile. There were 2,986 secondary classrooms with more than 5 students responding.

the range of 0.80 and above. They are also stable for a given teacher during the school year. (Corrected for measurement error, the correlations over time in classroom level responses in December and March of the same school year ranged between 0.70 and 0.85.)

Although we test below whether their judgments correspond with achievement gains, classrooms of students clearly differentiate among teachers. Tables 1 and 2 report the 25th and 75th percentiles of the classroom level agreement rates for the elementary and secondary items respectively. For instance, one of the questions asking students to provide feedback on a teacher's success at classroom management asks students to agree or disagree with the statement, "My classmates behave the way my teacher wants them to." In a quarter of classrooms, less than 23 percent of students agreed and in another quarter more than 53 percent of students agreed. In answering the question, "Our class stays busy and does not waste time", a quarter of classrooms had fewer than 44 percent of students agreeing and a quarter of classrooms had more than 71 percent of students agreeing.

Secondary school students seemed particularly willing to distinguish between teachers. Under the Tripod Challenge index for secondary school students, for example, students were asked to agree or disagree with the statement, "In this class, the teacher accepts nothing less than our full effort." In a quarter of classrooms, less than half of students agreed with that statement; in another quarter of classrooms, more than 81 percent of students agreed.

DESCRIPTIVE STATISTICS FOR THE SAMPLE

In Table 3, we report characteristics of the teachers and students who contributed data for this report, as well as the characteristics of the districts where they work and learn. Roughly four out of five of the teachers who volunteered are female, which reflects the gender mix of the districts where they teach. The racial and ethnic composition of the volunteer teachers also reflects the characteristics of the teachers in the districts where they work, with roughly a quarter of teachers being African American, and 7 to 9 percent being Latino. The main difference is that the volunteers are more likely to be young teachers, having worked at the districts for an average of 8 years, as compared to 10.2 years for teachers overall in their districts.

The students were also similar in terms of gender and race/ethnicity to the average student in their home districts. The main difference was that there was a slightly lower percentage of English language learners (15.2 as compared to 18.2) and special education students (11.3 as compared to 14.9) when compared to the districts from which they were drawn.

In Table 4, we report the sample means and distributions for each of the measures used in the study. (These data are reported at the course section level, not the teacher level. More than half of the teachers taught more than one course section.) For example, we calculated value-added on the state mathematics exam for 1531 course sections. (Because of the standardization of test scores to be mean zero, the mean for all of

Table 3. Comparison of Sample and Host Districts

TEACHER CHARACTERISTICS	SAMPLE	AVERAGE IN MET DISTRICTS
Percent Female	82.4	77.8
Average years employed by the district	8.0	10.2
Percent African American	25.2	24.7
Percent Latino/Latina	6.7	9.1
Percent White, Non-Hispanic	65.8	62.8
Percent Other Race/Ethnicity	2.3	3.4
STUDENT CHARACTERISTICS		
Percent Female	48.1	48.7
Percent African American	31.2	30.4
Percent Latino/Latina	36.0	33.9
Percent English language learners	15.2	18.2
Percent Special Education	11.3	14.9

Note: Sample of 4th through 8th grade students were drawn from New York City, Charlotte-Mecklenburg, Hillsborough (Florida), Dallas and Denver. The average in MET districts was weighted by the size of the sample in each district.

the value-added measures is near zero.) Moreover, the 10th percentile and 90th percentiles of section-level value-added in math implied that students lost .244 standard deviations and gained .266 standard deviations respectively relative to similar students with similar classmates elsewhere in each district. The total variance in estimated value-added on the state ELA tests was somewhat lower than we found in math, while the variance in estimated teacher effects on both the Balanced Assessment in Math (BAM) and the Stanford 9 Open-Ended was larger. (As we discuss below, the total variance in estimated value-added can be misleading because of measurement error. For instance, the variation in persistent teacher effects on the state ELA tests seems to be much smaller than that on the state math tests.)

In this report, we analyze student perception data for 2519 classrooms. When presented with each of the statements, students reported their level of agreement using a 5 point scale.[4] Following the methods used by the Tripod project in the past, we calculated the mean for each question attaching a value of 1 to 5 to each category of response. (For those questions that were negatively worded, the order was reversed.) The scores for each question were then standardized to have mean zero and standard deviation one.

4 On the secondary survey, the categories were labeled "totally untrue", "mostly untrue", "somewhat", "mostly true", "totally true". On the elementary survey, the 5 choices were "no, never", "mostly not", "maybe/sometimes", "mostly yes", "yes, always".

Table 4. The Sample Distribution for Each of the Measures

VARIABLE	FULL SAMPLE MEAN (S.D.) [N]	10TH PERCENTILE	90TH PERCENTILE	VARIABLE	FULL SAMPLE MEAN (S.D.) [N]	10TH PERCENTILE	90TH PERCENTILE
VALUE ADDED MEASURES:				**TRIPOD:**			
VA on State Math Test	0.005 (0.227) [1531]	-0.244	0.260	Sum of 7 C's	0.019 (0.444) [2519]	-0.547	0.551
VA on State ELA Test	0.001 (0.186) [1670]	-0.214	0.216	Control+Challenge	0.021 (0.566) [2519]	-0.709	0.735
VA on BAM Test	-0.005 (0.262) [1389]	-0.335	0.299	Other 5 C's	0.022 (0.542) [2519]	-0.673	0.664
VA on Stanford 9 OE ELA	-0.006 (0.345) [1533]	-0.397	0.389	Care	0.038 (0.512) [2519]	-0.629	0.672
				Control	0.027 (0.495) [2519]	-0.625	0.602
				Clarify	-0.009 (0.607) [2519]	-0.825	0.758
				Challenge	0.044 (0.483) [2519]	-0.575	0.631
				Captivate	0.011 (0.533) [2519]	-0.690	0.669
				Confer	-0.004 (0.507) [2519]	-0.661	0.593
				Consolidate	0.028 (0.479) [2518]	-0.579	0.605

Note: The sample size for each mean is reported inside the square brackets.

Analysis

Suppose you were a school leader trying to staff your school for a new school year. You would likely be asking yourself, "What does each teacher's past performance—in terms of student achievement, classroom observations, peer evaluations, etc.—say about how their students are likely to fare this year?" Every artifact of a teacher's practice—whether informal comments from parents and peers, direct classroom observations or (in an increasing number of school districts) the achievement gains of recent and past students—is potentially useful in answering that critical question. After all, the purpose of teacher evaluation is not to assess past performance for its own sake or to rehash past personnel decisions, but to inform professional development and staffing decisions going forward.

Our analysis plan mimics the school leader's question. We ask, *"How well do various aspects of a teacher's performance in one class or in one academic year help predict the student achievement gains in that teacher's classroom during another academic year or class?"* In this preliminary report, we test the predictive power—both individually and collectively—of student perceptions and available value-added data to predict a teacher's impact on students in another class or academic year.

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 estimates of value-added in *another* class.
- Second, focusing on those teachers for whom we have value-added estimates from a prior year (2008-09), we test whether measures of classroom practice in 2009-10 are related to the *past* value-added of each teacher.

If the measures are helpful in predicting performance in prior years and in other classes, they ought to be helpful in predicting a teacher's future impact on students.

HOW BIG ARE THE LONG-TERM DIFFERENCES BETWEEN TEACHERS?

As reported in Table 4, we see wide variation in student achievement gains between classrooms of students taught by different teachers in any given year. But does this reflect the effect of teachers—or simply random variation in student achievement in different classes? One way to resolve that question is to look at a given group of teachers and study their value-added with different groups of children. If the measures were purely random variation, we would not expect to see any systematic relationship between a teacher's value-added with *different* groups of children.[5] Accordingly, we focus on teachers for whom we have value-added estimates for *more than one group* of students—either in different sections of the same course or in different academic years. Three-fifths (58 percent) of the MET teachers in our sample taught more than one class of students in a given subject during the 2009-10 school year. For this group of teachers, we can look at the relationship

5 This is not a test for bias. A teacher could systematically be assigned the most well-behaved children year after year. Ultimately, we hope to resolve the bias question with the analysis of randomly assigned classrooms.

between the value-added in a given subject on a given test in one class and compare it to the value-added estimate for the same teacher in another class. Second, we estimated "value-added" during the year prior to the study (2008-09) for two-fifths (44 percent) of the MET teachers in this sample. (The latter information is only available for value-added on the state tests, since we did not administer the BAM or Stanford 9 OE tests during 2008-09.) Three quarters of the teachers in MET were in at least one of the two groups.

Table 5 reports estimates of the variation in teacher value-added, breaking the variance into two parts—that due to non-persistent sources of variation and that due to the persistent differences in teacher effects. We have two different ways of isolating the persistent and non-persistent components, using those teaching two sections of the same subject in 2009-10, as well as for those with value-added estimates in two years. The first column reports the total (unadjusted) variance in teacher value-added per course section that we observed, for each of the four tests. These estimates include both persistent (stable) differences in value-added as well as non-persistent differences between teachers. For many different reasons, value-added measures might fluctuate from year to year or from classroom to classroom. One reason is the natural variation that occurs when the identities of the students change from year to year. (This is analogous to the sampling variation associated with any random sample from a larger population.) When there are 20 students in an elementary classroom or 35 students in a secondary classroom in a given year, a few particularly talented or attentive youngsters in one year could lead to gains in one classroom that would be hard to replicate with another group. A second reason is any other non-persistent factor which influences a whole group of students simultaneously: a few rowdy kids who disrupt learning for everyone, a dog barking in the parking lot on the day of the test, a virulent flu outbreak the week of the state test, etc. A third reason is less than perfect test reliability. Any test covers only a sample of all the knowledge taught in a given year. Value-added could fluctuate simply because of the items used in any year and their inclusion or exclusion in the lessons taught by that teacher in that year.

Table 5. The Stable Component in Value-Added on Various Assessments

		DIFFERENT SECTION			PRIOR YEAR	
VARIABLE	TOTAL VARIANCE ONE SECTION	CORRELATION COEFFICIENT	IMPLIED VARIANCE OF STABLE COMPONENT	TOTAL VARIANCE PRIOR YEAR	CORRELATION COEFFICIENT	IMPLIED VARIANCE OF STABLE COMPONENT
TYPE OF TEST	[S.D. IN BRACKETS]		[S.D. IN BRACKETS]	[S.D. IN BRACKETS]		[S.D. IN BRACKETS]
State Math Test	0.053	0.380	0.020	0.040	0.404	0.016
	[0.231]		[0.143]	[0.20]		[0.127]
State ELA Test	0.032	0.179	0.006	0.028	0.195	0.005
	[0.178]		[0.075]	[0.166]		[0.073]
BAM Test	0.071	0.227	0.016			
	[0.266]		[0.127]			
Stanford 9 OE ELA	0.129	0.348	0.045			
	[0.359]		[0.212]			

Note: The standard deviation (s.d.) in value-added is the square root of the variance. The BAM scores and Stanford 9 scores were not available for any teacher in the year prior to the study.

When the between-section or between-year correlation in teacher value-added is below .5, the implication is that more than half of the observed variation is due to transitory effects rather than stable differences between teachers. That is the case for all of the measures of value-added we calculated. We observed the highest correlations in teacher value-added on the state math tests, with a between-section correlation of .38 and a between-year correlation of .40. The correlation in value-added on the open-ended version of the Stanford 9 was comparable, .35. However, the correlation in teacher value-added on the state ELA test was considerably lower—.18 between sections and .20 between years.

Does this mean that there are no persistent differences between teachers? Not at all. The correlations merely report the *proportion* of the variance that is due to persistent differences between teachers. Given that the total (unadjusted) variance in teacher value-added is quite large, the implied variance associated with persistent differences between teachers also turns out to be large, despite the low between-year and between-section correlations. For instance, the implied variance in the stable component of teacher value-added on the state math test is .020 using the between-section data and .016 using the between-year data. Recall that the value-added measures are all reported in terms of standard deviations in student achievement at the student level. Assuming that the distribution of teacher effects is "bell-shaped" (that is, a normal distribution), this means that if one could accurately identify the subset of teachers with value-added in the top quartile, they would raise achievement for the average student in their class by .18 standard deviations relative to those assigned to the median teacher.[6] Similarly, the worst quarter of teachers would lower achievement by .18 standard deviations. So the difference in average student achievement between having a top or bottom quartile teacher would be .36 standard deviations. That is far more than one-third of the black-white achievement gap in 4th and 8th grade as measured by the National Assessment of Educational Progress—closed in a single year!

The outcome with the smallest implied variance in teacher effects is value-added on the state ELA tests. This is a common finding in studies of this type across the country.[7] It is often interpreted as implying that teachers have less impact on students' reading and verbal skill. However, one possible explanation is the nature of the state ELA tests themselves. Most of the state ELA tests in our study (as in most states around the country) focus on reading comprehension, using short reading passages followed by multiple choice questions. However, outside the early elementary grades when students are first learning to read, teachers may have limited impacts on general reading comprehension.

There is some other evidence suggesting this may be the case. For instance, the regular version of the Stanford 9 test (not the open-ended version which we use in this study) uses a similar format to many of the state reading assessments. But unlike many state tests, the regular Stanford 9 has a "vertical scale," meaning that scores are intended to be comparable across grades. The original publishers of the Stanford 9 achievement test (Harcourt Educational Measurement, now owned by Pearson Education) administered their tests to a nationally representative sample of youth in April and October of 1995. If we are willing to assume that the differences between birth cohorts are small, we can take the growth in mean scores between the fall sample and spring sample for a given grade level as an estimate of the growth in achievement between fall and spring.

6 The mean value inside the top quartile of a normal distribution with a variance of .020 is .18.

7 For all 7 of the districts and states where Hanushek and Rivkin (2010) could find estimates of teacher effects on both math and reading, the variance in teacher effects on math was larger than that on reading.

Likewise, we can take the difference in scores between the spring sample in one grade and the fall sample in the next grade as approximating the amount of growth students achieve over the summer. For students in grades 1 through 3, the improvement in mean reading scores between October and April were larger than differences between April of one grade and October of the subsequent grade.[8] Because students generally spend more time in school between October and April than between April and October, such a finding implies youth are improving their reading comprehension *more* during the months when they are in school.

However, beginning in fourth grade, that is no longer true! The norm sample results imply that students improve their reading comprehension scores just as much (or more) between April and October as between October and April in the following grade. Scores may be rising as kids mature and get more practice outside of school. However, the above pattern implies that schooling itself may have little impact on standard reading comprehension assessments after 3rd grade.

But literacy involves more than reading comprehension. As the Common Core State Standards recently adopted in many states remind us, it includes writing as well. In fact, English teachers after grade 4 generally focus more on writing than teaching children to read. That is one of the reasons why we supplemented the state ELA tests by administering the Stanford 9 Open-Ended assessment, which provided students with reading passages and then asked students to provide written responses. The implied standard deviation in teacher effects on that alternative assessment Stanford 9 performance was .21, somewhat larger, in fact, than in math. In future analyses, we will be investigating whether teachers have a stronger influence on writing skills than they do on reading comprehension, by analyzing the writing prompts in some state assessments separately from the multiple choice reading comprehension questions.

However, one can never be certain that one is looking at a top-quartile teacher. To do so with anything approaching certainty would mean watching the teacher work with many thousands of children in many thousands of classrooms. Rather, we have to infer a teacher's effectiveness, based on their recent performance, classroom observations and student reports. Those inferences will naturally include some error. As a result, the difference between those who are *inferred to be* effective or ineffective will be smaller than the differences above. Yet, the better the predictors one has, the better that inference will be and the closer the evaluation system will come to discerning the large differences in effectiveness the measures suggest are there. In this report, we will be testing how large a difference one could infer with just two factors: value-added and student perceptions. In future reports, we will be adding other factors which could improve our inferences further, using classroom observations and the new teacher assessment.

How does the volatility in "value-added" compare to that of performance measures in other professions?

Quantitative indicators of performance are new to the education field. Understandably, many have raised questions about the statistical reliability of those measures. However, as it happens, the volatility in a teacher's value-added between years is no higher than for the performance measures used in Major League Baseball—a field known for its reliance on quantitative measurement. Smith and Schall (2000) studied batting averages and earned-run-averages (ERA) for major league baseball players. The between-season correlation in batting averages was .36. The between-season correlation for major league pitchers' ERA was even lower, .31. (Note that these estimates are lower than the correlation for math value-added, but higher than that found for teacher impacts on state ELA tests.) Despite such volatility, batting averages and ERA's are commonly used metrics for evaluating performance in baseball.

8 Harcourt Educational Measurement (1996), Tables N2 and N5.

DOES HIGH VALUE-ADDED COME AT THE EXPENSE OF CONCEPTUAL UNDERSTANDING?

As reported in Table 5, the correlation between a teacher's value-added on the state test and their value-added on the Balanced Assessment in Math was .377 in the same section and .161 between sections. Does this mean that the teachers who succeed in promoting gains on the state test are different from the teachers who are promoting gains on the more conceptually-challenging BAM test? No. Recall that both value-added measures (the BAM test, especially) are measured with error. Because of such measurement error, the correlation between *measured* value-added and *anything else* misstates the true correlation. (Indeed, for the *same* reason, the total, unadjusted variance in measured value-added overstates the variance in actual teacher effects. That's why we used only the smaller estimate of the variation in the persistent component of teacher effects, which is adjusted for measurement error.) In addition, to the extent that the two value-added measures are calculated for the same set of students, the correlation could be influenced by any common trait of students affecting both outcomes and not necessarily the teacher. (For this reason, the correlation in measured value-added on the two tests from different sections is biased downward, while the correlation in measured value-added using the two tests and the same group of students could be biased upward or downward.)

To calculate the true correlation in teacher effects on state math tests and the Balanced Assessment in Math, we make use of the fact that we measure each with two *different* groups of students. By studying whether those teachers who had high value-added on the state math test with one group of students and also tended to have high value-added on the BAM with another group of students, and by calculating the persistent variance in each of these measures using a calculation analogous to that we used in Table 5, we can gain some insight into whether those teachers who are successful in raising state math test scores also tend to be successful in promoting achievement on the Balanced Assessment. When we do that, we estimate the correlation between the persistent component of teacher impacts on the state test and on BAM is moderately large, .54.[9]

In other words, the teachers whose students show gains on the state tests also tend to see unusual gains on other tests. Because the BAM test focuses more on conceptual understanding and uses a very different format than most state tests, this would imply that those teachers who are showing strong value-added scores on the state test are not simply "teaching to the test". Their impact seems to generalize to other tests as well.

The correlation in the stable teacher component of ELA value-added and the Stanford 9 OE was lower, .37. However, one possible reason that the correlation was so much lower was a change in tests in NYC this year. When we exclude the observations from NYC, the estimated correlation in persistent teacher effects on state ELA tests and the Stanford 9 OE in reading was .59. We will be studying this issue further in future reports.

9 If the measured value-added for a given teacher on an assessment k and classroom j is $VA_k^j = \delta_k + \varepsilon_k^j$, then the correlation in teacher effects on the two types of tests can be written as $\rho_{\delta_{State},\delta_{Suppl}} = \frac{Cov(\delta_{State},\delta_{Suppl})}{\sqrt{Var(\delta_{State})Var(\delta_{Suppl})}}$. If the measurement error on each of the measures is independent of the true teacher effect on both measures, then we can estimate the correlation by dividing the between-section covariance between the two measures by the product of the standard deviation in the stable component in each measure. We estimate the latter using the covariance between-sections in each of the measures. The above correction is designed to adjust for any non-persistent factor included in the error, not simply sampling variation or the reliability of tests.

DO STUDENTS IN DIFFERENT CLASSROOMS SEE A TEACHER IN THE SAME LIGHT?

The usefulness of any potential predictor of effective teaching is related to its own stability across academic years and between different classrooms of students. While it is difficult enough to hit a moving target (such as predicting value added in another class or year) while standing on stable ground, it is even more difficult when the ground is bouncing around. If a given measure fluctuates from year-to-year or from class-to-class, its ability to predict any persistent difference in a teacher's effectiveness will be diminished.

Table 6 reports the between-section (same teacher) correlation for the student perception measures. In general, the student perception measures were highly correlated between sections taught by the same teacher. For instance, the between-section correlation for the overall composite measure—summing across the 7C's—was .67. Moreover, each of the 7C's showed similar degrees of consistency across classrooms, with between-section correlations ranging from .58 to .68.

Table 6. The Stability of Effectiveness Measures Between-Sections Taught by the Same Teacher, 2009–10

VARIABLE	CORRELATION [SAMPLE SIZE]	VARIABLE	CORRELATION [SAMPLE SIZE]
VALUE ADDED MEASURES:		TRIPOD:	
VA on State Math Test	0.381 *** [520]	Sum of 7 C's	0.668 *** [956]
VA on State ELA Test	0.180 *** [574]	Control+Challenge	0.601 *** [956]
VA on BAM Test	0.228 *** [452]	Other 5 C's	0.682 *** [956]
VA on Stanford 9 OE ELA	0.348 *** [514]	Care	0.669 *** [956]
		Control	0.657 *** [956]
		Clarify	0.557 *** [956]
		Challenge	0.642 *** [956]
		Captivate	0.685 *** [956]
		Confer	0.614 *** [956]
		Consolidate	0.648 *** [955]

Note: The sample size for each correlation is reported inside the square brackets. A *,**, or *** indicates a correlation that is significantly different from zero at the .10, .05 and .01 level respectively.

USING INDIVIDUAL MEASURES TO PREDICT VALUE-ADDED: MATHEMATICS

In Table 7, we report the (pair-wise) correlation between each of a set of measures and various measures of teacher value-added in math. In fact, we report the correlations with each of five different value-added measures: value-added on the state test from the *same section* in which the data were collected (column 1); value-added on the state test in a *different section* from the one where students were surveyed or instruction was observed (column 2); value-added on the state test in the *prior year* (column 3); value-added on the Balanced Assessment in Mathematics in the *same section* where students were surveyed or the observation was conducted (column 4) ; and value-added on the Balanced Assessment in Mathematics in a *different section* (column 5).

We report the correlations separately by same section, different section and different academic year for one reason: when two measures are drawn from the same group of students or same classroom, *both measures* are likely to share some common characteristic of a class of students which is not attributable to a teacher's

Table 7. Pairwise Correlations with Teacher Value-Added: Math

	VALUE-ADDED ON STATE MATHEMATICS TEST			VALUE ADDED ON BALANCED ASSESSMENT IN MATH		NUMBER OF TEACHERS
	SAME SECTION	DIFFERENT SECTION	PRIOR YEAR	SAME SECTION	DIFFERENT SECTION	
VA on State Math Test	1.000	0.380 ***	0.404 ***	0.377 ***	0.161 ***	1011
Disattenuated					0.542	
TRIPOD:						
Sum of 7 C's	0.212 ***	0.218 ***	0.203 ***	0.107 ***	0.114 ***	952
Disattenuated		0.433	0.346		0.296	
Care	0.158 ***	0.155 ***	0.146 ***	0.073 **	0.096 ***	952
Disattenuated		0.307	0.265		0.246	
Clarify	0.208 ***	0.237 ***	0.189 ***	0.093 ***	0.105 ***	952
Disattenuated		0.487	0.336		0.281	
Control	0.224 ***	0.171 ***	0.180 ***	0.182 ***	0.143 ***	952
Disattenuated		0.384	0.352		0.420	
Challenge	0.219 ***	0.216 ***	0.232 ***	0.080 **	0.115 ***	952
Disattenuated		0.436	0.404		0.301	
Captivate	0.158 ***	0.197 ***	0.152 ***	0.080 **	0.082 **	952
Disattenuated		0.388	0.258		0.210	
Confer	0.135 ***	0.166 ***	0.157 ***	0.049	0.091 ***	952
Disattenuated		0.336	0.275		0.241	
Consolidate	0.142 ***	0.181 ***	0.153 ***	0.052	0.050	952
Disattenuated		0.367	0.268		0.132	
Control+Challenge	0.256 ***	0.219 ***	0.235 ***	0.160 ***	0.149 ***	952
Disattenuated		0.465	0.435		0.414	
Other 5 C's	0.173 ***	0.201 ***	0.173 ***	0.075 **	0.091 ***	952
Disattenuated		0.395	0.298		0.234	

Note: A *,**, or *** indicates a correlation that is significantly different from zero at the .10, .05 and .01 level respectively. The correlation for "different section" was for at most one video observation in another section, so is likely to increase as more videos are scored. Disattenuated correlations under State Mathematics Test for different section and prior year are not reported as they are by definition 1.

practice. For example, even among those with similar prior academic achievement (which the value-added measures control for), one group of students may be unusually well behaved. In that section, a teacher's value-added is likely to be positive, but the students' perceptions of the classroom climate and the score of the video observation also may be unusually high. Therefore, if we were to focus solely on the same section correlations, we run the risk of overstating the predictive power of a given measure. In contrast, the different section or prior year correlations are based on data from distinct groups of students, where the teacher is the common factor.

Because the first row reports correlations for value-added estimates in math, the entry for the first column and first row is, of course, one. However, the value-added for a given teacher has approximately the same correlation with value-added from another section (.380) as it does with the previous year (.404). It appears that the stable component of a teacher's effect on students is shared to roughly the same degree between classrooms in the same year as between academic years.

Next, we report correlations for the student perception survey. The overall Tripod index was correlated to the same degree (.21 to .23) with all three value-added measures based on the state math test, whether within the same section, a different section or the prior year. The correlation of student perceptions with BAM value-added was lower (.11), although it remained true that the correlation with "same section" value-added was similar to the correlation with "different section" value-added.

All of the above correlations are unadjusted for measurement error. However, just as we did when estimating the correlation between teacher effects on state tests and supplemental tests, we can estimate the correlation in the underlying persistent traits, adjusted for measurement error. (These are reported in the table with the label "disattenuated", because measurement error typically leads to attenuated, or lowered, estimates of the relationships.) The disattenuated correlations between the Tripod index and math value-added in another section and another year was .43 and .34 respectively. The disattenuated correlation with value-added on BAM was only slightly lower, .30.

As reported in Table 7, the individual subscores of Tripod which were most strongly related to student achievement gains were "control" and "challenge", with an unadjusted correlation with teacher value-added in own section of .22. (The disattenuated correlations were .38 and .44 respectively.) As a result, we separate the Tripod index into two components: the first grouping consists of the "Control" and "Challenge" indices of the Tripod; the second grouping consists of the other five Tripod indices ("Care", "Clarify", "Captivate", "Confer" and "Consolidate"). Although both indices are positively related to value-added gains on both the state test and the BAM, the correlations were higher for the "Control" and "Challenge" indices than for the others. (See Tables 1 and 2 for a list of the questions underlying each of the indices.)

What is most important: an orderly environment, a caring teacher or lots of test preparation?

We studied the relationship between the mean student responses on each question on the Tripod survey for middle school students and the teacher's value-added in mathematics. The five questions with the strongest pair-wise correlation with teacher value-added were: "Students in this class treat the teacher with respect." (ρ=0.317), "My classmates behave the way my teacher wants them to."(ρ=0.286), "Our class stays busy and doesn't waste time." (ρ=0.284), "In this class, we learn a lot almost every day."(ρ=0.273), "In this class, we learn to correct our mistakes." (ρ=0.264) These questions were part of the "control" and "challenge" indices. We also asked students about the amount of test preparation they did in the class. Ironically, reported test preparation was among the weakest predictors of gains on the state tests: "We spend a lot of time in this class practicing for the state test." (ρ=0.195), "I have learned a lot this year about the state test." (ρ=0.143), "Getting ready for the state test takes a lot of time in our class." (ρ=0.103) Appendix Table 1 reports the correlations for each question.

USING INDIVIDUAL MEASURES TO PREDICT VALUE-ADDED: ENGLISH LANGUAGE ARTS

In Table 8, we report results for English Language Arts. Value-added on the state ELA exam in one section was positively correlated with value-added gains in another section taught by the same teacher, and value-added gains in the prior year. Indeed, as was true on the state math test, the between-section correlation (.18) was comparable to the between-year correlation (.20).

The overall index of the Seven C's from Tripod was positively correlated with all the ELA value-added measures, whether it was on the state test or on Stanford 9, in own section, other section or prior year. However, each of the correlations was lower than that found in math classrooms. Breaking the Tripod index into components, the combination of "control" and "challenge" was most strongly and consistently related to student achievement gains on all the above.

Table 8. Pairwise Correlations with Teacher Value-Added: ELA

	VALUE-ADDED ON STATE ENGLISH LANGUAGE ARTS			VALUE ADDED ON STANFORD 9 OE ELA		NUMBER OF TEACHERS
	SAME SECTION	DIFFERENT SECTION	PRIOR YEAR	SAME SECTION	DIFFERENT SECTION	
VA on State ELA Test	1.000	0.179 ***	0.195 ***	0.221 ***	0.093 ***	1096
Disattenuated					0.367	
TRIPOD:						
Sum of 7 C's	0.095 ***	0.070 **	0.099 ***	0.135 ***	0.063 **	1026
Disattenuated		0.195	0.250		0.121	
Care	0.029	0.027	0.039	0.081 **	0.002	1026
Disattenuated		0.075	0.105		0.004	
Clarify	0.087 ***	0.072 **	0.073 **	0.123 ***	0.064 **	1026
Disattenuated		0.198	0.186		0.121	
Control	0.142 ***	0.099 ***	0.084 **	0.158 ***	0.088 ***	1026
Disattenuated		0.294	0.236		0.183	
Challenge	0.128 ***	0.111 ***	0.162 ***	0.147 ***	0.092 ***	1026
Disattenuated		0.316	0.427		0.178	
Captivate	0.050	0.049	0.078 **	0.102 ***	0.041	1026
Disattenuated		0.136	0.196		0.080	
Confer	0.040	0.002	0.070 **	0.078 **	0.017	1026
Disattenuated		0.007	0.184		0.036	
Consolidate	0.080 **	0.067 **	0.092 ***	0.106 ***	0.079 **	1026
Disattenuated		0.189	0.238		0.154	
Control+Challenge	0.158 ***	0.118 ***	0.138 ***	0.180 ***	0.101 ***	1026
Disattenuated		0.342	0.374		0.203	
Other 5 C's	0.060 *	0.045	0.076 **	0.106 ***	0.042	1026
Disattenuated		0.124	0.194		0.080	

Note: A *,**, or *** indicates a correlation that is significantly different from zero at the .10, .05 and .01 level respectively. The correlation for "different section" was for at most one video observation in another section, so is likely to increase as more videos are scored.

COMBINING MEASURES TO PREDICT VALUE-ADDED IN ANOTHER SECTION: MATH AND ELA

In Table 9, we report on the predictive power of *combinations* of measures from one class section in forecasting value-added outcomes in another classroom of students taught by the same teacher in 2009-10.[10]

The first two columns in Table 9 present the mean value-added of the quarter of teachers with the most and least evidence of effectiveness in their other class. (Each of the rows in the table corresponds to a different type of information included in that evidence base.) The third column reports the difference in mean value-added between the top and bottom-ranked quarter of teachers.

Table 9. Predicting Value-Added in Another Section

DATA USED FOR ESTIMATING EFFECTIVENESS IN ORIGINAL SECTION	QUARTER WITH LEAST EVIDENCE OF EFFECTIVENESS	QUARTER WITH MOST EVIDENCE OF EFFECTIVENESS	DIFF BETWEEN TOP/BOTTOM 25%	
				(months)
OUTCOME: VALUE ADDED ON STATE MATH TEST				
Value-Added Only	-0.078	0.127	0.205 ***	7.390
Student Perceptions	-0.044	0.089	0.134 ***	4.816
Combining VA with Student Perceptions	-0.074	0.134	0.208 ***	7.494
OUTCOME: VALUE ADDED ON BAM				
Value-Added Only	-0.092	0.074	0.165 ***	5.956
Student Perceptions	-0.065	0.038	0.103 ***	3.715
Combining VA with Student Perceptions	-0.090	0.081	0.171 ***	6.149
OUTCOME: VALUE ADDED ON STATE ELA TEST				
Value-Added Only	-0.033	0.040	0.073 ***	2.633
Student Perceptions	-0.035	0.029	0.064 ***	2.311
Combining VA with Student Perceptions	-0.039	0.039	0.078 ***	2.818
OUTCOME: VALUE ADDED ON SAT9 TEST				
Value-Added Only	-0.178	0.151	0.329 ***	11.842
Student Perceptions	-0.044	0.037	0.081 **	2.929
Combining VA with Student Perceptions	-0.162	0.138	0.300 ***	10.784

Note: Since the quartiles were defined based on predictions from a regression that was fit to the value added data, conventional tests of the difference in value added between the quartiles tend to overstate the statistical significance. The p-values reported in this table adjust for this tendency. This was done by simulating the probability that the t-statistic testing the difference between quartiles would be greater than the observed t-statistic under the null that the variables being used to predict actually had no relationship to value added (so that there was no true difference between the quartiles). Monte Carlo experiments found that this method produced correct p-values.

10 Since the quartiles were defined based on predictions from a regression that was fit to the value added data, conventional tests of the difference in value added between the quartiles tend to overstate the statistical significance. The p-values reported in this table adjust for this tendency. This was done by simulating the probability that the t-statistic testing the difference between quartiles would be greater than the observed t-statistic under the null that the variables being used to predict actually had no relationship to value added (so that there was no true difference between the quartiles). Monte Carlo experiments found that this method produced correct p-values.

When using value-added on the state math test in the original section as the main source of evidence, the difference in actual student achievement gains in the other section between the top and bottom quartile of teachers was .21 student-level standard deviations. That's quite a large difference. It is between one-quarter and one-fifth of the black-white achievement gap closed in a single year. Nevertheless, it is smaller than the .36 standard deviation difference between top and bottom quartile teachers reported just a few pages above. Why? Is this inconsistent? Do teachers matter less than the evidence above would suggest? No. Recall that the .36 difference refers to the difference in student achievement for those teachers who *truly are* in the top versus bottom quartile of teacher effectiveness. The .21 standard deviation difference refers to the difference for those who are *inferred to be* in the top and bottom quartile, based on their recent performance (which is an imperfect indicator). As long as the evidence is imperfect, the latter difference must be smaller than the former difference.

However, it may be difficult for non-specialist readers to judge just how large this is. Another common way such effects are reported is in terms of "months of schooling". Such calculations typically require having a test with a vertical scale, meaning that the scores in different grades are comparable. There was no way to construct a common vertical scale using the various state tests. However, using the vertical scale scores from the Stanford 9 norm sample as well as age cut-offs for school enrollment in Los Angeles, Kane (2004) infers that 9 months of schooling is associated with a .25 standard deviation gain in performance.[11] Neal and Johnson (1996) use variation in educational attainment associated with quarter of birth and report that a year of schooling was associated with a .25 standard deviation gain on the Armed Forces Qualification Test. Although it's not ideal (and is likely to differ for different types of tests), we use that rule of thumb to convert from student-level standard deviation units into months of school. As reported in the last column of Table 9, a .21 standard deviation difference in scores would be roughly equal to 7.39 months of learning—in a 9 month school year!

The second row of Table 9 reports similar statistics if we were to use two groupings of the Tripod survey items as predictors—one pertaining to "control" plus "challenge" and a second capturing the other 5 C's. The difference in actual value-added between those with the strongest and weakest student perceptions was sizeable: .13 student-level standard deviations, or a half-year of schooling.

The third row combines the value-added and the student perception indices. The result is similar to the difference we saw with value-added alone: the difference between bottom and top quartile was .21 student standard deviations, roughly equivalent to 7.49 months of schooling in a 9-month school year.

Adding student perceptions on top of value-added increases the discernible difference in teacher effects from 7.39 to 7.49 months (which is statistically insignificant.) Does this mean that the student perception data are redundant because they add little predictive power? Not necessarily. Recall that the student perception data provides teachers with detailed data on the ways in which students perceive them. As long as they are positively related to student achievement gains (and they are), it would be worth including them in an effectiveness measure even if they do not substantially improve predictive power. The specificity of the feedback could justify their inclusion. Moreover, there are many grades and subjects where testing data and the value-added estimates they provide are not available.

11 We use this conversion factor for ease of explanation, fully recognizing that the actual factor is likely to vary by test, by grade level and by skill. For instance, learning gains in math and reading comprehension are smaller in later grades than early grades. However, if we had tests with a vertical scale in writing, we might see scores accelerate in later grades.

The next section of Table 9 reports the results of a similar exercise using teacher value-added on the BAM test. Combining the student perceptions and value-added on BAM as the predictors also reveals large differences in teacher effects (.17 standard deviations) between top and bottom quartile teachers.

Next, we report on a similar exercise with the state English Language Arts test. Combining value-added and student perception measures from another class, the difference in value-added for those teachers who were in the top and bottom quartile on the combined evidence measure was .078 student-level standard deviations.

Is this a large effect? Obviously, relative to the differences that are discernible in mathematics, it is not. However, recall that there are smaller persistent differences in teacher effects on state ELA tests to be "predicted." Relative to the differences in ELA value-added which are associated with other teacher traits, the effects are large. The difference between a top and bottom quarter of teacher is about twice the difference in value-added in ELA on the state tests associated with being a 3rd or 4th year teacher as opposed to a novice.[12] If a principal could identify a quarter of their novice teachers who are on average expected to generate reading gains on par with a 3rd or 4th year teacher, that would be a useful piece of information to have.

However, as we hypothesized above, the failure to discern large teacher effects may be a function of the limited nature of the state ELA tests, rather than a lack of teacher impact on literacy. In Table 9, we report a similar exercise using student scores on the Stanford 9 Open-Ended assessment. The difference in value-added between top and bottom quartile teachers is larger than reported for math—.300 student-level standard deviations.

PREDICTING VALUE-ADDED IN A PRIOR SCHOOL YEAR

Ultimately, our goal is to test the usefulness of teacher data from 2009-10 to anticipate differences in student achievement following random assignment in 2010-11. However, given that we are not even midway through the academic year, those results remain to be seen. Although we can't "predict" past value-added, we can "post-dict" a teacher's value-added in a prior school year. Specifically, we use performance data from 2009-10 to identify those teachers with large value-added in the prior year, 2008-09. Therefore, in Table 10, we use data from classrooms in 2009-10 to "predict" a teacher's value-added in the prior year.[13]

As reported in the top panel, when combining student perceptions and value-added scores from 2009-10, the difference in value-added in 2008-09 between those predicted to be in the top and bottom quartile based on that evidence was a sizeable .206 standard deviations or 7.4 months of schooling in a 9 month school year. Note that these differences are considerably larger than one would have been able to discern with student perceptions alone, where the difference for those with student perceptions in the top and bottom quartile was almost half as large, .129.

12 Many different studies (such as Kane, Rockoff and Staiger (2008) find that the value-added of the average 3rd or 4th year teacher is .06 to .09 student-level standard deviations above that of novice teachers.

13 If a teacher had more than one classroom in 2009-10, we calculated a simple average of the measures across all their available classrooms.

Table 10. Predicting Value-Added in a Prior Year

DATA USED FOR ESTIMATING EFFECTIVENESS IN ORIGINAL SECTION	QUARTER WITH WORST EVIDENCE OF EFFECTIVENESS	QUARTER WITH BEST EVIDENCE OF EFFECTIVENESS	DIFF BETWEEN TOP/BOTTOM 25%	
				(months)
OUTCOME: VALUE ADDED ON STATE MATH TEST				
Value-Added Only	-0.085	0.112	0.196 ***	7.070
Student Perceptions	-0.051	0.078	0.129 ***	4.636
Combining VA with Student Perceptions	-0.087	0.119	0.206 ***	7.430
OUTCOME: VALUE ADDED ON STATE ELA TEST				
Value-Added Only	-0.026	0.029	0.054 ***	1.960
Student Perceptions	-0.016	0.038	0.054 ***	1.928
Combining VA with Student Perceptions	-0.032	0.053	0.086 ***	3.078

Note: Since the quartiles were defined based on predictions from a regression that was fit to the value added data, conventional tests of the difference in value added between the quartiles tend to overstate the statistical significance. The p-values reported in this table adjust for this tendency. This was done by simulating the probability that the t-statistic testing the difference between quartiles would be greater than the observed t-statistic under the null that the variables being used to predict actually had no relationship to value added (so that there was no true difference between the quartiles). Monte Carlo experiments found that this method produced correct p-values.

As we learned when predicting student achievement in other sections taught by the same teacher, the predictive power of the teacher-level value-added estimates in English Language Arts is simply smaller. Even after combining value-added and student perceptions, the gap in prior year value-added between top and bottom quartile teachers was .086 student level standard deviations.

DIFFERING STAKES FOR STUDENTS AND TEACHERS

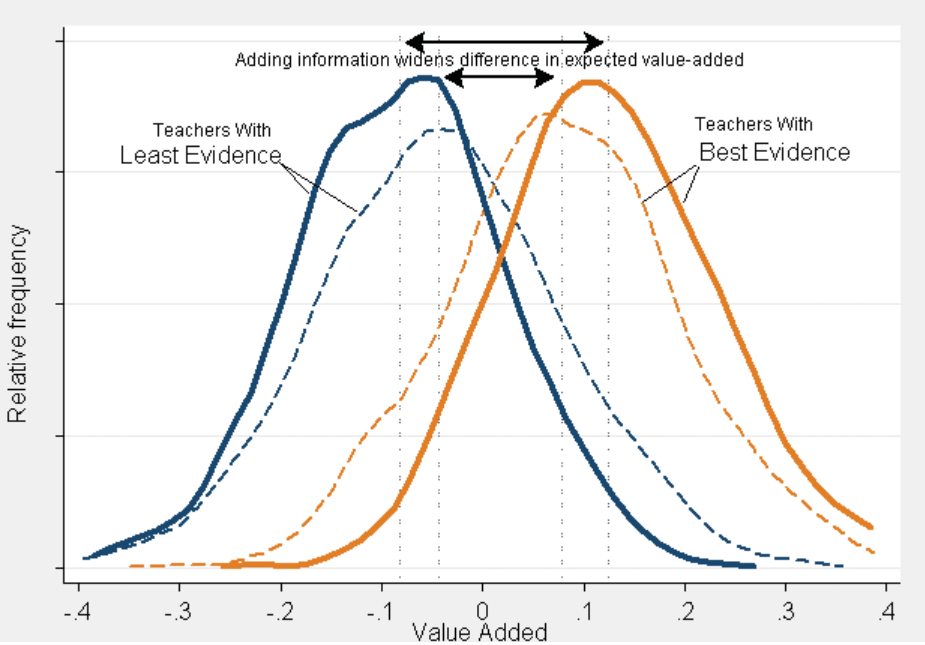
Suppose you were a school leader deciding whether to promote a beginning teacher to tenure after 2 or 3 years on the job. If your goal were to ensure strong student outcomes, you would ask two questions: "What is my best estimate—incorporating all I know about this teacher—of their effectiveness?" and "Is my best estimate of their effectiveness higher or lower than the average novice teacher I could recruit?" If your best estimate is that a given teacher is *more* effective than the average novice teacher you could recruit, then you should do what you can to retain the teacher, since doing so will increase expected student achievement. However, if the teacher is worse than the average novice, then you can raise expected student achievement by not tenuring the teacher. As harsh as that seems, it would be even more harsh not to do so, given their expected impact on students.

One's "best estimate" is likely to be the *mean* effectiveness of teachers with similar track records, similar relationships to students, similar teaching styles. It must be admitted that such evidence is not perfect. It cannot be. Some of the teachers with a given track record will turn out to be *more* effective than one thought, and some teachers will turn out to be *less* effective. Inevitably, some decisions based on such data will turn out to be mistaken. But that's the predicament we face with virtually every decision we make; we often have to make decisions with the imperfect evidence we have. Typically, we'd want school leaders to decide based on the best available evidence of a teacher's effectiveness.

Given the stakes involved in a tenure decision, *any* information in a teacher's portfolio which leads a principal to increase (or decrease) their estimate of a teacher's effectiveness is potentially valuable, because it gives a principal an opportunity to raise student achievement with a more discerning decision. Consider Figure 1. The orange lines depict the distribution of value-added of the quarter of teachers with the best evidence of effectiveness in 2009-10, while the blue lines represent the distribution in value-added for those with the least evidence of effectiveness. The dotted lines depict the distributions of value added when that evidence is limited to student perceptions; the solid lines depict the distributions when value-added data from 2009-10 is combined with student perceptions in 2009-10. When more information is added, two things happen. First, the two distributions are farther apart. In fact, the *difference* in mean effectiveness between the quarter of teachers with the best and worst evidence almost doubles from .129 to .206 when better information is added. Second, the distribution of true effects within each group narrows. Because the mean difference widens and because the tails of each distribution narrows, *the degree of overlap between the two groups lessens.*

The discussion above runs contrary to the usual narrative about value-added—which implies that using the data is a risky gambit, prone to mistakes. Why is the intuition of so many so wrong? The main reason is that many people simply forget all the bad decisions being made now, when there is essentially no evidence base available by which to judge performance. Every day, effective teachers are being treated as if they were the same as ineffective teachers and ineffective teachers are automatically granted tenure after two or three years on the job. Given that we know there are large differences in teacher effects on children, we are effectively mis-categorizing *everyone* when we treat everyone the same. Value-added data adds information. Better information will lead to fewer mistakes, not more. Better information will also allow schools to make decisions which will lead to higher student achievement.

Figure 1. Adding Value-Added Widens the Difference in Expected Effectiveness and Reduces Overlap



Note: The figure reports the distribution of value-added scores for the quarter of teachers with the most and least evidence of effectiveness. The dotted lines refer to the distributions based only on student perceptions. The solid lines report the distributions when value-added data from another section are added.

Conclusion

The evidence of wide differences in student achievement gains in different teachers' classrooms is like a colossal divining rod, pointing at the ground, saying, "Dig here." Dig here if you want to learn what great teaching looks like. Dig here if you want to better understand what teachers do to help students learn. This is where you will learn about ways to generate dramatically different results for kids.

With the support of the Bill & Melinda Gates Foundation, we have begun to dig. Although we have only begun to scratch the surface, the results so far are encouraging. Two types of evidence—student achievement gains and student feedback—do seem to point in the same direction, with teachers performing better on one measure tending to perform better on the other measures. In other words, it is possible to combine in a coherent package a teacher's student achievement results with feedback on specific strengths and weaknesses in their practice. We will be adding other measures—such as classroom observations and new teacher assessments—in future reports.

The public debate over measuring teacher effectiveness usually portrays only two options: the status quo (where there is no meaningful feedback for teachers) and a seemingly extreme world where tests scores alone determine a teacher's fate. Our results suggest that's a false choice.

Reinventing the way we evaluate and develop teachers will eventually require new infrastructure, perhaps using digital video to connect teachers with instructional coaches, supervisors and their peers. However, 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 (which could reach virtually every classroom, including those in non-tested grades and subjects);
- retraining principals and instructional coaches to do classroom observations in a more meaningful way; and
- delivering such data in a timely way to school principals and teachers.

These are all fairly low-cost ways to get started (especially important in this time of austerity). However, just as we have tried to do in this report, states and districts need to be disciplined enough to regularly check—in those classrooms where student achievement measures are available along with the other aspects of the evaluation, such as classroom observations and student perceptions—that the collection of measures they assemble allows them to "explain" some minimum amount of the variation in student achievement gains between teachers and that the measures 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 is to verify that—on average—those who shine in their evaluations are producing larger student achievement gains. For instance, a state or school district could replicate (annually, for instance) the results in Tables 9 and 10 with their own data—using evaluation results in one course section or academic year to predict value-added in another grade or academic year—to ensure that their feedback systems remain on track.

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, refine our approach to classroom observation—to find those teachers who truly excel, support them and develop others to generate similar results.

Technical Appendix

In order to generate value-added estimates for a teacher for each type of test, we first standardized the scores at the student level to have mean 0 and standard deviation 1 within each district, year and grade level. We then estimated the following equation using student level data:

$$S_{it} = X_{it}\beta + \overline{X}_{jkt}\gamma + \theta S_{it-1} + \lambda \overline{S}_{jkt-1} + \varepsilon_{it}$$

where the i subscript represent the student, j subscript represents the teacher, the k subscript represents the particular course section, the t subscript represents the year, X is a vector of student characteristics including race, gender, free or reduced price lunch status, ELL status, participation in gifted and talented programs, $\overline{X}_{ikt}$ represents the mean of these student characteristics by class, S_{it-1} represents student baseline scores and $\overline{S}_{jkt-1}$ represents mean student baseline scores in the class. (We estimated separate specifications for each district and grade level.) To generate teacher-level value-added estimates ($\hat{\tau}^S_{jkt}$) for the test S, we averaged the residuals from the above equation by teacher, section and year. This is similar to a random effects specification. (In other work, we have found such a specification to be largely equivalent a teacher fixed effects specification, since the lion's share of the variation in student characteristics is within classroom, as opposed to between classrooms.)

We suppose that the value-added estimate is composed of two components, a stable component (representing the "true" teacher effect for a given type of test, S), τ^S, and all other non-persistent classroom shocks and sampling variation, η^S_{jkt} .

$$\hat{\tau}^S_{jkt} = \tau^S + \eta^S_{jkt}$$

For a given teacher and section, the non-persistent component (η^S_{jkt}) is likely to be correlated across different tests. However, we are assuming that the η^S_{jkt} component is not correlated between sections for the same teacher or between years.

With the above set up, we can estimate the portion of the variation in $\hat{\tau}^S_{jkt}$ that is "stable" by studying the correlation within teacher across sections in a given year or between years. Moreover, we can calculate the correlation in the "stable" components of the teacher effects (τ^S) for two tests by taking the covariance in $\hat{\tau}^S_{jkt}$ for different tests across different sections and dividing it by the square root of the product of the standard deviations in τ^S for each of the tests. (The latter are estimated by the covariance within teacher, within test, across sections).

When calculating the standard errors on the difference between top and bottom quartile teachers in Tables 8 and 9, we realize that the typical standard errors generated by OLS are likely to be understated—because the initial regressions using covariates to predict value-added differences was fit against the same value-added estimates. As a result, the standard errors used for the last column were generated using bootstrap techniques.

Appendix Table 1

Pairwise Correlations with Math Value Added (Middle School)

CARE	
My teacher in this class makes me feel that s/he really cares about me.	0.228
My teacher seems to know if something is bothering me.	0.153
My teacher really tries to understand how students feel about things.	0.193
CONTROL	
Student behavior in this class is under control.	0.243
I hate the way that students behave in this class.	-0.176
Student behavior in this class makes the teacher angry.	-0.223
Student behavior in this class is a problem.	-0.242
My classmates behave the way my teacher wants them to.	0.286
Students in this class treat the teacher with respect.	0.317
Our class stays busy and doesn't waste time.	0.284
CLARIFY	
If you don't understand something, my teacher explains it another way.	0.220
My teacher knows when the class understands, and when we do not.	0.218
When s/he is teaching us, my teacher thinks we understand even when we don't.	-0.174
My teacher has several good ways to explain each topic that we cover in this class.	0.244
My teacher explains difficult things clearly.	0.250
CHALLENGE	
My teacher asks questions to be sure we are following along when s/he is teaching.	0.198
My teacher asks students to explain more about answers they give.	0.222
In this class, my teacher accepts nothing less than our full effort.	0.214
My teacher doesn't let people give up when the work gets hard.	0.240
My teacher wants us to use our thinking skills, not just memorize things.	0.202
My teacher wants me to explain my answers—why I think what I think.	0.194
In this class, we learn a lot almost every day.	0.273
In this class, we learn to correct our mistakes.	0.264
CAPTIVATE	
This class does not keep my attention—I get bored.	-0.215
My teacher makes learning enjoyable.	0.224
My teacher makes lessons interesting.	0.229
I like the ways we learn in this class.	0.242
CONFER	
My teacher wants us to share our thoughts.	0.177
Students get to decide how activities are done in this class.	0.173
My teacher gives us time to explain our ideas.	0.170
Students speak up and share their ideas about class work.	0.217
My teacher respects my ideas and suggestions.	0.207
CONSOLIDATE	
My teacher takes the time to summarize what we learn each day.	0.189
My teacher checks to make sure we understand what s/he is teaching us.	0.246
We get helpful comments to let us know what we did wrong on assignments.	0.203
The comments that I get on my work in this class help me understand how to improve.	0.226
TESTPREP	
We spend a lot of time in this class practicing for [the state test].	0.195
I have learned a lot this year about [the state test] .	0.143
Getting ready for [the state test] takes a lot of time in our class.	0.103

Note: The correlations were calculated using classroom level means (the weighted average of possible responses from 1 through 5, with 5 indicating strong agreement) with middle school math value-added.

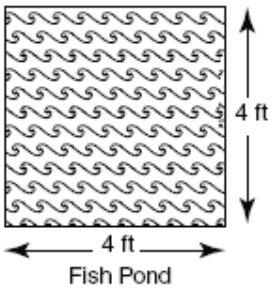
Appendix 1

Sample 8th Grade BAM Item[14]

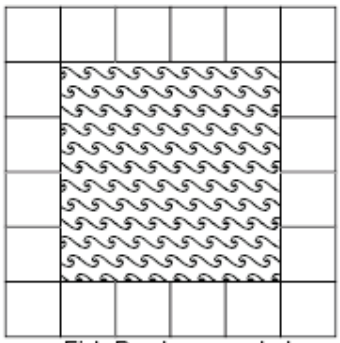
Fish Ponds

This problem gives you the chance to:
- find a number pattern in real spatial context and express the rule
- extend the rule to two variables

.



Fish Pond


Fish Pond surrounded by paving stones

Chris works at a garden center that sells square fish ponds and paving stones.

The paving stones are squares with sides one foot long.

1. Use the diagram above to figure out how many paving stones are needed to surround a fish pond that is 4 feet by 4 feet. _____

2. Chris begins to make a table to show how many paving stones are needed to surround square ponds of different sizes. Fill in the empty boxes in the table.

Side of pond in feet	1	2	3	4	5
Number of paving stones	8				

14 Copyright Mathematics Assessment Resource Service, 2001.

3. How many paving stones are needed to surround a fish pond that is 20 feet by 20 feet? Explain how you figured it out.

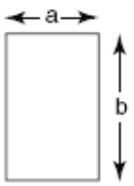
4. Chris has 48 paving stones. Find the size of the largest square pond the paving stones can surround. Explain how you figured it out.

5. The garden center sells many different sizes of square fish ponds.

Write down a rule that will help Chris figure out how many paving stones are needed to surround square ponds of different sizes.

6. The garden center decides to sell rectangular ponds.

Find a rule that will help Chris figure out how many paving stones are needed to surround rectangular ponds of different sizes.



Appendix 2

Example from Stanford 9 Open-Ended Reading Assessment[15]

Open-ended
Reading Sample-Intermediate Level

STANFORD 9
ACHIEVEMENT TEST SERIES
NINTH EDITION

Carl's Discovery by Sharon Phillips Denslow

Carl was helping his father pull weeds in the yard when he found a toad sitting in a hole in the middle of the yard. Carl tried gently pulling it out of its hole, but the toad swelled itself up and dug in its back legs.

"It's a little early yet for the toad to come out," Carl's father said.

It was warmer the next morning, and Carl found the toad sitting beside its hole. Just before dark, Carl checked on the toad again. It was back inside the hole.

The next morning when Carl went outside, frost had turned the grass into tiny feather icicles. The toad was snug in its hole.

"You're a pretty smart toad," said Carl. A chilly wind blew for two days. Carl put a curved white seashell in front of the toad's hole to keep the wind from whistling down it.

By the next weekend, the grass was scraggly enough for Carl's father to get out the lawn mower.

"Why doesn't the toad leave the hole?" Carl asked his father.

"It's warm enough now."

"It's still cold at night," answered his father.

Gradually, the ground grew warmer, and spring flowers began to bloom. One day Carl went barefoot for the first time. He noticed bugs flying and buzzing around the flowers. "You have something to eat now," he told the toad. The next morning Carl looked for the toad and finally found it at the edge of the garden, in the shade of a young tomato plant. Carl grinned. Near the big toad sat several small toads no bigger than Carl's fingernail.

Sharon Phillips Denslow has written several books for children, including Riding with Aunt Lucy. *She also works in a library.*

Get the Big Picture

How would you describe Carl to someone who had not read this story? Use details from the story to support your ideas.

Take a Closer Look

Carl said, "You're a pretty smart toad." Why did he say that?

Be a Critic

Do you think the author of this story knows much about toads?

Why do you think that?

87

15 Copyright 1996 by Harcourt.

References

Allen, J. P., Gregory, A., Mikami, A. Y., Lun, J., Hamre, B., & Pianta, R. C., (2010). Observations of effective teaching in secondary school classrooms: Predicting student achievement with the CLASS-S. Manuscript submitted for publication.

Curby, T. W., Stuhlman, M., Grimm, K. J., Mashburn, A. J., Chomat-Mooney, L., Downer, J., Hamre, B. K., & Pianta, R. C. (in press). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal.*

Danielson, C. (2007) *Enhancing Professional Practice: A framework for teaching* Alexandria, VA: Association for Supervision and Curriculum Development.

Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. (2010). Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms. Manuscript submitted for publication.

Hamre, B. K. & Pianta, R. C. (2007). Learning opportunities in pre-school and early elementary classrooms. In R. Pianta, M. Cox and K. Snow (Eds.) *School readiness and the transition to kindergarten in the era of accountability* (pp. 49-84). Baltimore, MD: Paul H. Brookes.

Hanushek, E.A., and Rivkin, S.G. (2010). Generalizations About Using Value-Added Measures of Teacher Quality. *American Economic Review, 100* (2): 267–271.

Harcourt Educational Measurement (1996) *Stanford Achievement Test Series, Ninth Edition, Technical Data Report,* (San Antonio, TX: Harcourt).

Kane, T.J. (2004) "The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations", Working Paper, (New York: William T. Grant Foundation). Available at http://www.wtgrantfoundation.org/publications_and_reports/browse_reports/kane_working_paper.

Kane, T.J. & Staiger, D.O. (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation" *NBER Working Paper* No. 14607.

Kane, T. J., Rockoff J. & Staiger,D.O. (2008) "What Does Certification Tell Us about Teacher Effectiveness?: Evidence from New York City" *Economics of Education Review 27*(6), 615-31.

La Paro, K.M., Pianta, R.C., & Stuhlman, M. (2004). Classroom Assessment Scoring System™ (CLASS™): Findings from the pre-k year. *Elementary School Journal, 104*(5), 409-426.

Malmberg, L-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (in press). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology.*

Marder, M., Walkington, C., Abraham, L., Allen, K., Arora, P., Daniels, M., Dickinson, G., Ekberg, D., Gordon, J., Ihorn, S. & Walker, M. (2010). *The UTeach Observation Protocol (UTOP) Training Guide* (adapted for video observation ratings). UTeach Natural Sciences, University of Texas Austin.

Neal, D.A., Johnson, W.R. (1996) "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy, 104*(5), pp. 869-95.

Nye, B., Konstantopoulos, S. and Hedges, L. (2004) "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis (26)*3, 237-257.Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R. & Morrison, F. (2008). Classroom effects on children's achievement trajectories in elementary school. American *Educational Research Journal, 45*(2), 365-397.

Pianta, R.C. & Hamre, B.K. (2010). *Overview of the Classroom Assessment Scoring System (CLASS).* Presentation at the annual meeting of the American Educational Research Association, Denver, CO.

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R.C., Hamre, B. K., Haynes, N. J., Mintz, S. L. & L Paro, K. M. (2009). *Classroom Assessment Scoring System (CLASS), Secondary Manual.* Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS).* Baltimore, MD: Paul H. Brookes.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*(3), 225-238.

Pianta, R.C., Mashburn, A. J., Downer, J. T., Hamre, B. K. & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23,* 431-451.

Raver, C.C., Jones, S.M., Li-Grining, C.P., Metzger, M., Smallwood, K., Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly, 23*(1) 10-26.

Rothstein, J. (2010) "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement" *Quarterly Journal of Economics (125)*1, 175-214.

Smith G. & Schall, T. (2000). Do baseball players regress toward the mean? *The American Statistician, 54,* 231-245.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org