MET project

RESEARCH PAPER

Have We Identified Effective Teachers?

Validating Measures of Effective Teaching Using Random Assignment

Thomas J. Kane
Daniel F. McCaffrey
Trey Miller
Douglas O. Staiger

BILL & MELINDA GATES foundation

ABOUT THIS REPORT: This report presents an in-depth discussion of the technical methods, findings, and implications of the Measures of Effective Teaching (MET) project's random assignment study of teaching effectiveness measures. A non-technical summary of the analysis is in the policy and practitioner brief, *Ensuring Fair and Reliable Measures of Effective Teaching.* All MET project papers and briefs may be found at www.metproject.org.

ABOUT THE MET PROJECT: The MET project is a research partnership of academics, teachers, and education organizations committed to investigating better ways to identify and develop effective teaching. Funding is provided by the Bill & Melinda Gates Foundation.

Partners include representatives of the following institutions and organizations: American Institutes for Research, Cambridge Education, University of Chicago, The Danielson Group, Dartmouth University, Educational Testing Service, Empirical Education, Harvard University, National Board for Professional Teaching Standards, National Math and Science Initiative, New Teacher Center, University of Michigan, RAND, Rutgers University, University of Southern California, Stanford University, Teachscape, University of Texas, University of Virginia, University of Washington, and Westat.

ACKNOWLEDGMENTS: The Bill & Melinda Gates Foundation provided support for the project. Steve Cantrell and Sarah Buhayar from the Bill & Melinda Gates foundation played key roles in designing and implementing the MET project. Kerri Kerr provided excellent project management. Ann Haas, Scott Naftel, and Marian Oshirio at RAND created the analysis files and generated the value-added estimates used in our analyses. Rita Karam, Karin Kitchens, Louis Mariano, Kata Mihaly, John Pane, and Andie Phillips at RAND and Lisa Gardner, Carol Morin, Robert Patchen, and Alka Pateriya at Westat contributed tremendously to the work on randomly assigning rosters to the MET project teachers. Alejandro Ganimian provided research assistance at Harvard. We received help from many dedicated staff in six MET project partner school districts: the Charlotte-Mecklenburg Schools, the Dallas Independent Schools, the Denver Public Schools, the Hillsborough County (Fla.) Public Schools, the Memphis Public Schools, and the New York City Schools.

In addition to MET project partners who reviewed early drafts of this report, we would like to thank the following external experts who read and provided written feedback: Anthony Bryk at the Carnegie Foundation for the Advancement of Teaching, Dan Goldhaber at the University of Washington, Sean Reardon at Stanford University, and Jonah Rockoff at Columbia University. The lead authors accept full responsibility for any remaining errors in the analysis.[1]

January 2013

1 The lead authors and their affiliations are Thomas J. Kane, Professor of Education and Economics at the Harvard Graduate School of Education and principal investigator of the Measures of Effective Teaching (MET) project; Daniel F. McCaffrey, Senior Statistician at the RAND Corporation; Trey Miller, Associate Economist at the RAND Corporation; and Douglas O. Staiger, Professor of Economics at Dartmouth College.

Table of Contents

I. Introduction

To develop, reward, and retain great teachers, school systems first must know how to identify them. We designed the Measures of Effective Teaching (MET) project to test replicable methods for identifying effective teachers. In past reports, we described three approaches to measuring different aspects of teaching: student surveys, classroom observations, and a teacher's track record of student achievement gains on state tests (Kane and Staiger, 2010 & 2012). In those analyses, we could only test each measure's ability to predict student achievement gains non-experimentally, using statistical methods to control for student background differences. For this report, we put the measures to a more definitive and final test. First, we used the data collected during 2009–10 to build a composite measure of teaching effectiveness, combining all three measures to predict a teacher's impact on another group of students. Then, during 2010–11, we randomly assigned a classroom of students to each teacher and tracked his or her students' achievement. We compared the predicted student outcomes to the actual differences that emerged by the end of the 2010–11 academic year.[1]

Simply naming the key dimensions of teaching and measuring them are difficult enough. The task of validating such measures is complicated by the systematic sorting of students to teachers.[2] Within the schools in our sample, we saw considerable differences in the students assigned to different teachers, in terms of prior test scores, race, and ethnicity. We can control for the student characteristics that we observe. However, students might differ in ways that are invisible to us. If the same unmeasured student background traits lead to more significant student achievement gains, so that some teachers appear more effective than they truly are, we could be fooled into thinking we have measured teaching, when all we've done is identified teachers whose students were exceptional in some unmeasured way.

Therefore, without the extraordinary step of random assignment, it would be impossible to know if the measures that seemed to be related to student achievement gains in our past reports are truly identifying better teaching.

By randomly assigning students to teachers, we made it very unlikely that the students assigned to seemingly more or less effective teachers would be different in measured or unmeasured ways. Therefore, following random assignment, we studied the achievement of students assigned to teachers with differing prior measures of effectiveness. Looking across all the sample teachers and their predicted outcomes, we asked two questions: (1) Did the measures of effective teaching successfully identify sets of teachers who produced higher student achievement gains on average? And (2) did the *magnitude* of the differences correspond with what we would have predicted based on their measured effectiveness in 2009–10?

Here's what we found: First, the measures of effectiveness from the 2009–10 school year did identify teachers who produced higher average student achievement following random assignment. As a group, the teachers identified as more effective produced greater student achievement growth than other teachers in the same school, grade, and subject. Second, the magnitude of the achievement gains they generated was consistent with our expectations. In other words, the measures of effectiveness—even though they were collected before

1 The analyses in this paper are based on data from six MET project partner districts: Charlotte-Mecklenburg, Dallas, Denver, Hillsborough County (Fla.), New York City, and Memphis.

2 That is, students assigned to different teachers differ more in terms of background characteristics than would be expected under random assignment.

random assignment, when classrooms of students were assigned to teachers in the usual manner—generated predictions of teachers' impact on students that were borne out when teachers were subsequently randomly assigned.

Figures 1 and **2** provide a visual summary. We calculated teachers' predicted effectiveness based on the 2009–10 measures.[3] We then sorted teachers into 20 equally sized groups based on their predicted impact on students. (There were about 40 teachers in each group.) We then calculated the mean achievement at the end of 2010–11 for the students randomly assigned to them. The vertical axis measures end of year achievement on state tests following random assignment. The horizontal axis measures predicted student achievement. We plotted outcomes for each of the 20 groups. (Units are student-level standard deviations.)

The dashed line represents the relationship we would have seen if the actual difference in achievement equaled the predicted difference in achievement (slope = 1). For both math and English language arts (ELA), we see that the groups of teachers with higher predicted impacts on student achievement had higher actual impacts on student achievement following random assignment. Moreover, the actual impacts were approximately in line with the predicted impacts in math and ELA.[4]

For a subsample of students, we also studied teacher impacts on outcomes other than state tests.[5] We find that students randomly assigned to those teachers judged to be more effective on state tests also scored better on these other assessments. In math, the Balanced Assessment in Mathematics was designed to probe students' conceptual understanding. In English, the open-ended reading test we used required students to read a series of passages and write short-answer responses to prompts testing their comprehension. When a teacher was predicted to improve student achievement on state tests by one standard deviation, his or her students' performance on the supplemental assessments increased by .7 standard deviations on average. His or her students were also .84 standard deviations more likely to self-report that they enjoyed being in the class.

To guard against over-interpretation, we add two caveats: First, a prediction can be correct on average but still be subject to prediction error. For example, many of the classrooms taught by teachers in the bottom decile in the measures of effectiveness saw large gains in achievement. In fact, some bottom decile teachers saw average student gains larger than those for teachers with higher measures of effectiveness. But there were also teachers in the bottom decile who did worse than the measures predicted they would. Anyone using these measures for high stakes decisions should be cognizant of the possibility of error for individual teachers.

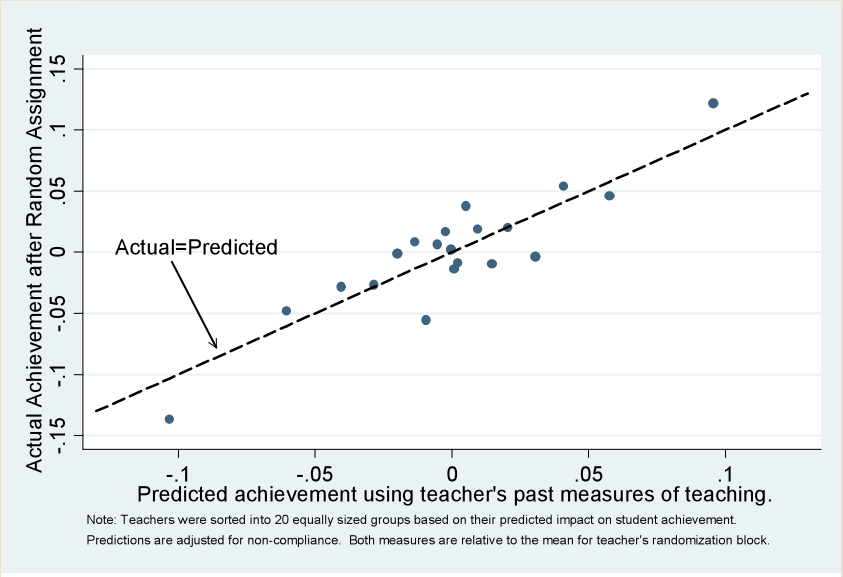
3 The mean end-of-year achievement was first adjusted for students' baseline test scores and demographics. The predicted achievement is the result of our first-stage estimate described later in the text, and as such it is corrected for test volatility and non-compliance with the randomization. Since teachers were randomly assigned within a given school, grade, and subject, we calculated both measures relative to the mean in their randomization block.

4 The differences in Figures 1 and 2 are smaller than the differences reported in earlier MET project reports. Due to non-compliance, only about 30 percent of the randomly assigned difference in teacher effectiveness translated into differences in the effectiveness of students' actual teacher. If all the students had remained with their randomly assigned teacher, we would have predicted impacts roughly three times as big. Our results imply that, without non-compliance, we would have expected to see impacts just as large as included in earlier reports.

5 We observed students' performance on the state test as long as they remained in the school district and took the test with the same identification code. State test scores were available for 86 percent of the randomly assigned sample in grades 4 through 8. In contrast, we collected the additional outcome data for students who remained in one of the MET project classrooms following randomization.

Figure 1

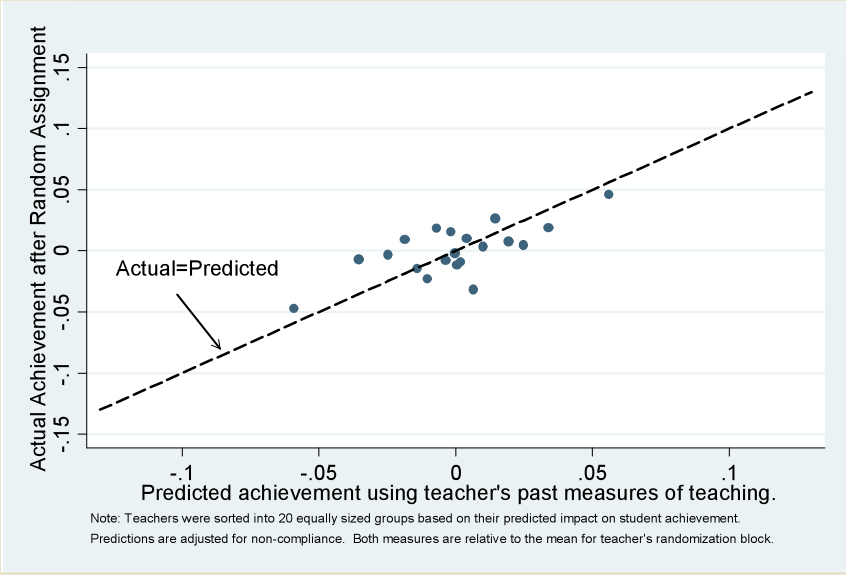
ACTUAL AND PREDICTED ACHIEVEMENT OF RANDOMIZED CLASSROOMS (MATH)



Note: Teachers were sorted into 20 equally sized groups based on their predicted impact on student achievement. Predictions are adjusted for non-compliance. Both measures are relative to the mean for teacher's randomization block.

Figure 2

ACTUAL AND PREDICTED ACHIEVEMENT OF RANDOMIZED CLASSROOMS (ENGLISH LANGUAGE ARTS)



Note: Teachers were sorted into 20 equally sized groups based on their predicted impact on student achievement. Predictions are adjusted for non-compliance. Both measures are relative to the mean for teacher's randomization block.

Second, as a practical matter, we could not randomly assign students or teachers to a different school site. As a result, our study does not allow us to investigate the validity of the measures of effectiveness for gauging differences across schools.[6] The process of student sorting across schools could be different than sorting between classrooms in the same school. Yet school systems attribute the same importance to differences in teachers' measured effectiveness between schools as to within-school differences. Unfortunately, our evidence does not inform the between-school comparisons in any way.

The remainder of this report is organized as follows:

In Section II, we describe the methods for generating the measures of effectiveness, which are essentially predictions of a teacher's impact on another group of students. We describe how we used the data collected during 2009–10 to predict the magnitude of a teacher's impact in 2010–11.

In Section III, we describe the statistical model we used to test for the impact of teachers with different measures of effectiveness. One of the challenges we faced was adjusting for non-compliance, when students and teachers ended up in classrooms other than the ones to which they were randomly assigned. We describe the statistical technique we used to infer the impact of students' actual teachers, despite non-compliance.

In Section IV, we compare the sample of students and teachers to others in the six districts. Among other differences, we describe the degree of teacher-student sorting in the years prior to randomization. Before randomization, both the MET project sample and the other classrooms systematically sorted students to teachers. Year after year, some teachers consistently received the highest- or the lowest-scoring students in their schools. It is clear that even the MET project sample schools were not randomly assigning students to teachers prior to 2010–11.

In Section V, we study the data for signs of problems that could invalidate our test. We check if random assignment yielded comparable groups of students across all levels of the teacher's measured effectiveness; we check whether sample attrition was related to the assigned teacher's measure of effectiveness; and we study whether the experiment may have been contaminated by student exposure to non-randomized peers. We did not see evidence that our test was invalid.

In Section VI, we present the results of the analysis. We study impacts of teacher ratings on state test scores, on tests other than the state tests, and on several other student reported measures. We report impacts by grade level. We report the results for high school students separately. We also compare the predictive validity of a variety of different value-added models. Each of the value-added models tries to approximate the causal impacts of teachers following random assignment, but some seem to do a better job than others.

We conclude in **Section VII.**

6 Although they rely on natural movements of teachers between schools (as opposed to randomly assigned transfers), Chetty, Friedman, and Rockoff (2011) do not find evidence of bias in estimated teacher effectiveness between schools.

II. Predicting the Magnitude of a Teacher's Future Students' Achievement Gains

During 2009–10, the MET project collected a variety of measures of teaching practice: We measured students' achievement gains, we collected four videos of classroom practice and scored them, and we asked students in those classes about specific aspects of their teachers' practice.[7] Because students had been assigned to teachers in the usual manner (that is, non-randomly), we adjusted the measures for a list of student traits that we could measure.

STUDENT ACHIEVEMENT GAINS AND VALUE-ADDED MODEL

Specifically, to infer a teacher's track record of student achievement impacts, we started by estimating the following value-added model, using students' achievement on state tests, S_{it}, as the dependent variable:

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

where the *i* subscript represents the student, the *j* subscript represents the teacher and the *t* subscript represents the year. X_{it} is a vector of student characteristics including race, gender, free or reduced-price lunch status,[8] English language learner (ELL) status, and participation in gifted and talented programs. $\overline{X}_{jt}$ represents the mean of the student characteristics, prior achievement, and free or reduced-price lunch status, for the students taught by a given teacher in a given year. The latter is intended to capture "peer effects" on an individual student's achievement. S_{it-1} represents the individual student's baseline scores, and $\overline{S}_{jt-1}$ represents a teacher's mean student baseline score in that year (and is also intended to capture the effects of peers).

Equation (1) describes our primary specification for estimating teacher effects on student achievement. Later in the paper, we test a variety of alternative value-added models, excluding peer effects (γ=0 and λ=0), excluding controls for prior student achievement (θ=0), and excluding controls for student demographic characteristics and program participation (β=0).[9]

Given that the state tests in each of the six MET project districts were different and measured on different scales, we first standardized the test scores to be normally distributed, with mean zero and standard deviation of one by district, grade level, subject, and calendar year.[10] We estimated equation (1) separately for each district and grade.

7 Although most of the concern regarding bias has focused on the achievement gain measures, the non-test-based components could also be biased by the same unmeasured student traits that would cause bias with the test-based measures.

8 We do not have free and reduced-price lunch status data for students in Charlotte-Mecklenburg Schools.

9 We also tested the predictive power of observations and student surveys on their own.

10 We standardized test scores using a rank-based standardization method or van der Waerden scores (Conover, 1999), which first ranked students based on the original test score and then assigned a standardized score based on the average score for students with that rank if the underlying scores were standard normal (giving students with the same score the average across all the associated ranks).

The residuals from the model above represent the degree to which each student outperformed or underperformed similarly situated students. By averaging the residuals across a teacher's students, we used these to generate teacher-level value-added estimates ($\hat{\tau}^{S}_{jt}$) for a teacher in each subject.[11]

ADJUSTED OBSERVATION SCORES AND STUDENT SURVEY RESPONSES

Although the term "value-added" is typically used when student achievement on end-of-year state tests is the measure to be adjusted, better-prepared students could also influence a teacher's classroom observations or student surveys. Therefore, we used student characteristics and baseline test scores and their classroom means to adjust a teacher's classroom observation scores and student surveys as well.[12] In practice, most districts and states do not adjust their classroom observations or student survey results for the student baseline characteristics. Yet we found they were highly correlated. For student surveys, the correlation between the unadjusted means and those adjusting for student baseline scores and peer characteristics was .92. For classroom observations, the correlation between adjusted and unadjusted was .95.

PREDICTING A TEACHER'S IMPACT ON STUDENTS — SCALING FOR ERROR

Even if it is an unbiased measure, the value-added or adjusted student achievement gain from any single year, $\hat{\tau}^{S}_{jt}$, is an imperfect measure of a teacher's impact. Because it contains error, the variation will be wider than it would be in the true measure. When there is a lot of error, we should be less willing to "go out on a limb" with large positive or large negative predictions. We should rein in the estimates, shrinking both the large positive impacts and the large negative impacts back toward zero. This idea of pulling estimates back toward zero to reflect the variance in underlying effectiveness is often referred to as "shrinkage" or "Empirical Bayes estimation" and there is a large literature describing its properties.

However, the challenge with shrinkage estimation is determining the appropriate shrinkage factor to use. We estimate the shrinkage factor by studying the degree to which the measure fluctuates from year to year. If a teacher's true, underlying effectiveness does not change from one year to the next, and if the errors in measured effectiveness are independent across years (e.g., based on different samples of students), then the stability (or lack of volatility) of measured effectiveness indicates the amount of error and, correspondingly, the amount of shrinkage required. When the measure does not fluctuate much, it implies there is less error and less need for shrinkage; when there is considerable fluctuation from year to year, it implies more error and more need for shrinkage.

11 This is essentially a random effects specification. In unpublished work, we have found such a specification yielded value-added estimates that are highly correlated to those from a teacher fixed effects specification, since the lion's share of the variation in student characteristics is within classroom as opposed to between classrooms (Ehlert et al., under review, report similar results with data in Missouri). This is particularly true in our specification, which controls for between-classroom variation in student characteristics with section averages (Mundlak, 1978).

12 For the student survey score, we estimated models identical to equation (1) but used the standardized student-survey score in place of the standardized test score as the dependent variable. For the classroom observation score (for which there are no student-level data), we used the residuals from the following equation, where M_{jkt} is a measure of a teacher's classroom observation score:

$$M_{jkt} = \overline{X}_{jkt}\gamma + \lambda\overline{S}_{jkt-1} + u_{jkt}$$

Therefore, to estimate the shrinkage factor, we used the value-added measure from one year to predict value-added in a different year. Specifically, if we had value-added for two years (year t and year t') we could estimate the following equation by linear regression:

$$(2)\ \hat{\tau}^{S}_{jt} = \beta_o + \beta_1 \hat{\tau}^{S}_{jt} + \eta_{jt}$$

Using equation (2), a teacher's predicted value-added, $\hat{\beta}_o + \hat{\beta}_1 \hat{\tau}^{S}_{jt}$, consists of three components: (i) the teacher's average adjusted student achievement gain during school year t ($\hat{\tau}^{S}_{jt}$), (ii) the coefficient on the achievement gain measure, $\hat{\beta}_1$, and (iii) the estimated intercept, $\hat{\beta}_o$. The estimated slope parameter, $\hat{\beta}_1$, takes on a value between 0 and 1. The more the measure fluctuates from year to year, the smaller $\hat{\beta}_1$ will be, the more error we will infer, and the more we will shrink the measure from one year to predict a teacher's impact on students in another year.

But we have more than one type of measure at our disposal in order to predict a teacher's impact in another year. To incorporate the additional measures, we extended the approach above, simply adding more predictor variables to equation (2) when these additional measures were available.[13] In addition to value-added from 2009–10, we used the average of a teacher's scores for four videos that were scored on Charlotte Danielson's Framework for Teaching (Danielson, 1996) and the average student responses from students surveyed using the Tripod measure, developed by Ron Ferguson at Harvard University (Ferguson, 2009).[14]

To estimate the best linear combination of the measures for predicting a teacher's impact in another year, we did not use 2010–11 data as the outcome data—since those are the outcomes we will be using to test the predictions later. Instead, we used value-added estimates for the 2008–09 school year as the outcome variable:

$$(3)\ \hat{\tau}^{S}_{j2008-09} = \beta_o + \beta_1 \hat{\tau}^{S}_{j2009-10} + \beta_2 FFT^{adj}_{j2009-10} + \beta_3 Tripod^{adj}_{j2009-10} + \beta_4 Z_j + \eta_{jt}$$

where $FFT^{adj}_{j2009-10}$ represents a teacher's score on domains 2 and 3 of Charlotte Danielson's Framework for Teaching and $Tripod^{adj}_{j2009-10}$ represents a teacher's average score across all seven domains of the Tripod student survey.[15] The *adj* superscript is meant to suggest that both were adjusted for classroom mean test scores and demographics. We also include commonly observed teacher characteristics (Z_j) as additional predictors: experience (0–1, 2–3, or 4+ years) and an indicator for whether a teacher has a master's degree.

This approach takes account of the errors in each measure and generalizes shrinkage to include multiple component measures. In a companion report, Mihaly et al. (2013) provide more details on this approach to forming a composite or combined measure of effectiveness.

13 More precisely, for each teacher, we estimated a regression of the form given in equation (3) that included only the variables that were available for that teacher, and we used the resulting coefficients to form the prediction of that teacher's effectiveness. Thus, for teachers not in the MET project, we did not include FFT and Tripod but had the student achievement gain measures and teacher background variables. In addition, predictions for 2011 incremented each teacher's experience accordingly.

14 The student questionnaire is available at www.metproject.org.

15 As described in Kane and Staiger (2012), the project collected teacher scores on a number of other instruments as well: the Classroom Assessment Scoring System (CLASS), the Protocol for Language Arts Teacher Observations (PLATO), the Mathematical Quality of Instruction (MQI), and the UTeach Teacher Observation Protocol (UTOP) . However, since one of our goals was to approximate what a given district or state could do with its data, we used only one instrument, Framework for Teaching, in this analysis.

Obviously, there are unmeasured student characteristics that influence students' achievement in each year that we did not control for statistically. But it is only when these are *persistently* associated with some teachers (i.e., there are unmeasured student characteristics that affect a teacher's value-added estimate year after year), that the predictions will be systematically wrong. In estimating equations (2) and (3), we would interpret the persistent sorting as lower volatility. Following the logic above, we would overstate the proportion of variance due to actual teacher effectiveness and understate the shrinkage. However, as long as teachers do not get students with these same unmeasured characteristics year after year, then non-persistent fluctuations in classroom composition for individual teachers would be treated as another source of random error and the predictions would be scaled accordingly.

Table 1 reports estimates of the coefficients for equations (2) and (3). The equations were estimated separately in elementary grades (grades 4 and 5 in the top panel) and middle school grades (grades 6 through 8 in the bottom panel). Columns 1 and 5 contain results for equation (2). For neither math nor ELA and for neither elementary nor middle schools would we conclude that the value-added measures were equivalent to a random number. In each case, the coefficient on the value-added estimate in one year when predicting value-added in another year is statistically different from zero. The results imply that we would multiply a teacher's value-added scores in math by .396 and .512 in elementary and middle school, respectively, when trying to predict their results in another year. In other words, for each student-level standard deviation a teacher generates with his or her students this year, we would expect differences 40 to 50 percent as large next year. Middle school teachers tend to specialize in math, have more students with math scores, and therefore have less error.

The degree of volatility is larger in English: The coefficients on 2009–10 value-added were .350 and .117 in elementary and middle school, respectively. In other words, we would place less weight on value-added on state ELA tests when predicting a teacher's future achievement gains in English. (In the 2010 MET project report, we speculated that the lower reliability of the state ELA value-added gains was due to the reliance on multiple choice reading comprehension questions in the state ELA tests, which do not capture teacher effects on student writing.)

Columns 2 and 6 of Table 1 report the parameter estimates for equation (3), adding student surveys and classroom observations to the mix. In all the specifications, when value-added on state tests are available, they received the greatest weight in the composite measure of effectiveness. However, in elementary ELA, the coefficient on the student survey measure was also statistically significant. In middle school ELA, the coefficient on teacher observations was statistically significant.

Columns 3, 4, 7, and 8 report the results of using the student survey and classroom observations on their own to predict a teacher's student achievement gains. When used alone, each of these measures has a statistically significant coefficient in math. For elementary ELA, the observations were not statistically significantly predictive by themselves. For middle school ELA, the student surveys were not statistically significantly predictive, while the observations were. Later in the paper, we test the ability to predict a teacher's outcomes when using all of the measures together as well as when using each of the measures alone.

In our subsequent analyses, our combined measure of effectiveness for teacher j following random assignment, $\hat{\hat{\tau}}^{S}_{j}$, is the predicted value from equation (3). It combines evidence from value-added estimates from 2009–10, $\hat{\tau}^{S}_{j2009-10}$, with classroom observation scores on the Danielson instrument, $FFT^{adj}_{j2009-10}$, evidence from student surveys, $Tripod^{adj}_{j2009-10}$ (also from that year), and teacher background variables, as available.

Table 1

USING TEACHER PERFORMANCE MEASURES IN 2009–10 TO PREDICT STUDENT ACHIEVEMENT GROWTH IN 2008–09

Elementary Grades

	State Math				State English Language Arts			
	1	2	3	4	5	6	7	8
Value-Added Measure from 2009–10	0.396***	0.410***			0.350***	0.306***		
	(0.029)	(0.041)			(0.031)	(0.040)		
Student Survey Score (Tripod) in 2009–10		0.060	0.164***			0.074*	0.147***	
		(0.051)	(0.053)			(0.040)	(0.040)	
Classroom Observation Score (FFT) in 2009–10		0.042		0.110*		0.013		0.050
		(0.054)		(0.058)		(0.042)		(0.044)
Controls for Teacher Experience, MA Degree?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	782	377	405	392	828	390	417	403
R-squared	0.203	0.240	0.032	0.024	0.134	0.169	0.038	0.012

Middle School Grades

	State Math				State English Language Arts			
	1	2	3	4	5	6	7	8
Value-Added Measure from 2009–10	0.512***	0.526***			0.117***	0.299***		
	(0.036)	(0.052)			(0.0440)	(0.061)		
Student Survey Score (Tripod) in 2009–10		0.035	0.084**			-0.014	0.010	
		(0.033)	(0.033)			(0.029)	(0.027)	
Classroom Observation Score (FFT) in 2009–10		-0.005		0.089**		0.060**		0.077***
		(0.034)		(0.034)		(0.027)		(0.024)
Controls for Teacher Experience, MA Degree?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	559	270	312	303	586	277	316	303
R-squared	0.273	0.318	0.033	0.069	0.026	0.142	0.03	0.069

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all teachers of MET project students (including non-MET project teachers) with non-missing values on the outcome and predictor variables. The FFT and Tripod measures are only available for MET project teachers. The Tripod student survey used a five-point scale, which was standardized at the student level. The FFT score is on a scale from one to four. The coefficients reflect the specification in Equations 2 and 3 in the text. Standard errors are robust to the presence of heteroskedasticity.

In combining those components, we weighted each component by the coefficients reported in Table 1, $\hat{\beta}_o$ through $\hat{\beta}_4$. The more volatile the variable, $\hat{\tau}^S_{jt}$, the closer all these coefficients will be to zero. As such, $\hat{\tau}^S_j$ not only combines evidence from three different types of measures (value-added, classroom observations, and student surveys), it also adjusts for the year-to-year volatility in the outcome measure.

III. Testing the Causal Effect of a One-Unit Change in Measured Effectiveness

We used the 2009–10 school year data, when students were assigned to teachers non-randomly, to predict which teachers would be more and less effective with another group of students. Then we randomly assigned classrooms to teachers in 2010–11. We asked, *"Does a one-unit difference between two teachers in their measure of effectiveness translate into a one-unit difference in their students' achievement?"* If the predicted impacts on students reflect the causal effects of individual teachers in a given school and grade, then the average difference in student outcomes following randomization should correspond with the predicted differences. If not, we should see a different relationship between the predictions and student outcomes.

RANDOM ASSIGNMENT

In spring and summer 2010, the project team worked with staff at participating schools to identify the teachers who met the necessary requirements for random assignment. To be eligible for randomization, two things had to be true: (1) teachers had to be part of a group of two or more MET project teachers scheduled to teach the same grade and subject, and (2) their principals had to view the teachers as being capable of teaching any of the rosters of students designated for the group of teachers. We call these groups of teachers "randomization blocks." Schools then constructed rosters of students for the randomization blocks and submitted them to the study team. Within each randomization block, analysts at the RAND Corporation randomly assigned each of the rosters to the participating MET project teachers. (The randomization process is described in more detail in Appendix A.)

During the 2010–11 school year, we monitored the randomly assigned students and the teachers who taught them. Ideally, every randomly assigned student would have been taught by the teacher to whom the roster was assigned. But we could not force students, teachers, or principals to comply. The assignments had been made in summer 2010, before schools were certain which students or teachers were going to appear when school opened. In the weeks and months following random assignment, some students transferred to other schools or to other teachers' classes in the same school; some teachers left teaching or taught different course sections or grades than planned. In other cases, schools simply did not implement the randomization. (We know this because the students assigned to one teacher would all end up in another teacher's classroom, and some schools informed us that they would not implement the random assignments.) In the end, many students ended up with an "actual" teacher who was different from their "assigned" teacher.

Nevertheless, we were able to exploit the random assignment to infer the effect of the "actual" teacher, using a two-step process known as "instrumental variables" or "IV estimation." Essentially, we asked two questions, both of which relied exclusively on the randomly assigned rosters to estimate a causal effect: First, we asked how much of a difference it made for students' subsequent achievement to be assigned to an effective teacher's

classroom (whether or not they remained there)? Second, we asked how much of an impact being assigned to an effective teacher's classroom had on the measured effectiveness of one's actual teacher (which was driven by the extent of compliance)? Below, we explain how we generated these two sets of estimates and combined them to infer the impact of one's actual teacher.[16]

INSTRUMENTAL VARIABLES ESTIMATION FOR DATA WITH NON-COMPLIANCE

Our goal is to estimate the coefficient on the measured effectiveness of actual teachers, γ_1, in the following equation:

$$(4)\ S_{i2011} = \gamma_1^{\hat{\tau}^S_{jActual}} + \phi_k + \varepsilon_i$$

where S_{i2011} is the student's achievement at the end of the 2010–11 school year (following random assignment), $\hat{\tau}^S_{jActual}$ is the measured effectiveness of a student's actual teacher in 2010–11, and ϕ_k is a fixed effect for the randomization block (usually a given school, grade, and subject). The coefficient γ_1 captures how much student test scores increase for each unit difference in the measured effectiveness of their actual teacher. If the measured effectiveness ($\hat{\tau}^S_j$) correctly predicts a teacher's effectiveness on average, then the coefficient, γ_1, should equal one.[17, 18]

But while we could be confident that the effectiveness of the randomly assigned teacher was not correlated with unmeasured student characteristics, the actual teacher's effectiveness could well be related to unmeasured student characteristics, given that any reshuffling following random assignment was likely not to be random. Presumably, the same processes of sorting that occur in the typical year were at work in the months following random assignment. Therefore, to estimate γ_1, we first estimated how much of the difference in the assigned teacher's effectiveness ($\hat{\tau}^S_{jAssigned}$) translated into differences in the actual teacher's measure of effectiveness ($\hat{\tau}^S_{jActual}$). In the context of instrumental variable estimation, this step is commonly referred to as the "first stage" model. To the extent that some students moved schools or classrooms and some principals ignored the random assignment rosters, a one-unit difference in assigned teacher's measured effectiveness would translate into a less than one-unit difference in the actual teacher's measure of effectiveness.

16 We use a statistical technique known as "instrumental variables," in which the effectiveness of the assigned teacher is an instrumental variable, an exogenous determinant of the effectiveness of the actual teacher. Although the effectiveness of a student's actual teacher, $\hat{\tau}^S_{jActual}$, is not randomly assigned and, therefore, could be correlated with unmeasured determinants of end-of-year achievement, ε_i, the randomly assigned teacher's effectiveness, $\hat{\tau}^S_{jAssigned}$, should be unrelated to ε_i. As long as the only route through which $\hat{\tau}^S_{jAssigned}$ affects student outcomes is through its effect on actual teacher effectiveness, $\hat{\tau}^S_{jActual}$, then the instrumental variables estimator should generate a consistent estimator of the effect of actual teacher effectiveness, γ_1.

17 If γ_1 diverges from one, this is evidence that the predictions are biased in the sense that $E\left(S_{i2011} \mid \hat{\tau}^S_{jActual}\right) \neq \hat{\tau}^S_{jActual}$. This property of an estimator is referred to as "unbiased prediction" in the economics literature or "calibrated prediction" in the statistics and forecasting literature. That is, the predictions systematically overstate or understate differences between teachers in their impacts on student test scores (perhaps because of the types of students they teach).

18 Model (4) includes only the teacher's measured effectiveness and a control for the randomization block. However, it is common in experimental studies to include additional covariates to improve precision (Snedecor and Cochoran, 1989). Accordingly, our preferred model adds to Model (4) student background characteristics as covariates (including the student's prior year test scores, program participation, and demographics).

Next, we studied the degree to which the *assigned* teacher's measured effectiveness ($\hat{\bar{\tau}}^{S}_{jAssigned}$) affected student achievement. Note this value will not equal γ_1 because some students were not taught by the assigned teacher. In instrumental variables estimation, this is commonly referred to as the "reduced form" or "intent-to-treat" model.

The instrumental variables technique estimates the effect of a student's actual teacher indirectly, by essentially taking the ratio of these two effects: the effect of the assigned teacher's measure of effectiveness on student achievement divided by the effect of the assigned teacher's effectiveness rating on the measure of effectiveness for the actual teacher:

$$\hat{\gamma}_1^{IV} \cong \frac{Effect\ of\ Randomly\ Assigned\ Teacher\ Rating\ on\ Student\ Achievement}{Effect\ of\ Randomly\ Assigned\ Teacher\ Rating\ on\ Actual\ Teacher\ Rating}$$

Both the numerator and denominator are estimated using the randomly assigned teacher's measured effectiveness. Therefore, both effects can be estimated without contamination by non-random student sorting. (That is, we estimate the causal effects of the randomly assigned teacher's measured effectiveness on the student's actual teacher rating and on the student's achievement.) As long as the only way in which the assigned teacher's effectiveness influences student achievement is through the actual teacher's effectiveness, the above should yield a consistent estimate of the coefficient on the measured effectiveness of a student's actual teacher on achievement, γ_1.[19]

As we report in the next section, compliance varied considerably by district and by school. As a result, we included in the instrumental variables interactions between assigned teacher effectiveness and district, school, or randomization block identifiers. The point estimates were very similar, with and without these interactions, although the inclusion of the interactions allowed for more precision.[20]

COMPLIANCE RATES

Table 2 reports the proportion of students remaining with their randomly assigned teacher through the end of the 2010–11 school year, as well as the proportion moving to another teacher in the same randomization block, the proportion remaining in the school in a classroom outside of the randomization block, the proportion attending another school in the district, and the proportion with no data on teacher assignments (who are likely to have moved outside the district).

The compliance rates are reported by district. The three districts with the highest compliance rates were Dallas, Charlotte-Mecklenburg, and Hillsborough, where 66, 63, and 56 percent of the students, respectively, remained with their randomized teacher throughout the school year. In those districts, the most common form of non-compliance was to move to a classroom outside of the randomization block, rather than move to another classroom in the randomization block.

19 The instrumental variable estimator generates an estimate of the Local Average Treatment Effect (LATE). In other words, we are estimating the incremental impact of teacher effectiveness for those classrooms that complied.

20 We estimated the model using a LIML version of the IV estimator. In our setting, LIML is the optimal (maximum likelihood) estimator if we consider each school (or randomization block) as a separate experiment in which the compliance rate varied randomly across the experiments (Chamberlain and Imbens, 2004). Moreover, the statistical properties of LIML estimates are superior to other IV estimators such as two-stage least squares in a setting such as ours, with many weak instruments (Stock, Wright, and Yogo, 2002).

Table 2

COMPLIANCE WITH RANDOMIZED TEACHER ASSIGNMENTS

	Remaining in Randomization Block					
Proportions	**Same Teacher**	**Different Teacher**	**Remaining in School**	**Other Schools in District**	**Missing**	**Total**
4th–8th Grade Sample by District (Math & ELA stacked):						
Dallas	0.656	0.025	0.253	0.010	0.055	1.00
Charlotte-Mecklenburg	0.634	0.140	0.177	0.009	0.040	1.00
Hillsborough	0.560	0.163	0.236	0.010	0.031	1.00
New York	0.452	0.189	0.209	0.009	0.141	1.00
Denver	0.395	0.242	0.241	0.000	0.122	1.00
Memphis	0.274	0.230	0.325	0.012	0.159	1.00
High School Sample:						
Math, ELA, and Biology	0.452	0.137	0.303	0.036	0.072	1.00

Note: The 4th to 8th grade sample constitutes the main randomized sample. The high school sample includes all 9th grade students who were assigned to a randomized teacher. Those in the "missing" column had teacher IDs corresponding to a teacher no longer teaching in the district.

The three districts with the lowest compliance rates were New York, Denver, and Memphis, where 45, 40, and 27 percent of the students remained with their randomized teacher, respectively. Students in these districts were more likely to attend class with a different teacher inside the randomization block. They were also more likely to be missing data on their teacher assignments. (For more on the randomization process, see Appendix A.)

IV. Teacher and Student Characteristics

Table 3 reports the characteristics of three different groups of teachers and the students in their classrooms: (1) the 1,181 teachers who were randomized; (2) the 3,802 other teachers in the MET project schools who were not randomized; (3) the 17,153 teachers from the same districts working in schools without a MET project teacher. All teachers were teaching math or English in grades 4 through 8.

The MET project teachers and principals provided their consent to participate. As a result, it is not a random sample of teachers or schools. Nevertheless, in 2009–10, the mean and standard deviation in their value-added scores in 2009–10 were similar to those of teachers who chose not to participate. (As in the earlier MET project reports, we calculated the "signal variance" in teachers' true impacts on student achievement by taking the covariance in estimated value-added in 2008–09 and 2009–10.) The implied standard deviation in "true teacher effects" ranged from .16 to .18 in math and from .10 to .12 in ELA.

However, while the MET project teachers' value-added scores may have been similar to their colleagues', the mean baseline scores of the students of the randomized MET project teachers were higher in both math and ELA than those assigned to other teachers. During the 2009–10 school year, the randomized MET project teachers taught students with prior year performance .14 standard deviations higher in math and ELA than other teachers in the same schools (.101 versus -.040 in math and .089 versus -.048 in ELA). The gap between the students of randomized MET project teachers and the students in non-MET project schools was similar, suggesting that the randomized teachers were working with somewhat more advanced students on average in their own schools and relative to those in non-MET project schools.[21]

We use as an indicator of student sorting the standard deviation in the mean baseline scores of students assigned to different teachers in a given year. We measure persistent sorting using the covariance in a teacher's students' baseline test scores over time.[22] If a teacher gets a group of high-scoring students one year but not the next, that may or may not have occurred by chance that year, but it is not persistent sorting. Despite having higher scores on average, there was considerable evidence of sorting based on prior student achievement among the MET project teachers in the years before random assignment. For instance, the standard deviation in a teacher's mean student baseline math scores was .486 among the MET project randomized teachers from within the same school, grade, and subject. While that was somewhat less than the sorting among the non-randomized teachers inside MET project schools or among teachers in the non-MET project schools (which was .590 and .652, respectively), this means that the standard deviation in mean baseline performance between classrooms was nearly half as large as the standard deviation in scores between individual students across the district (since scores have been standardized to have a standard deviation of one at the student level).

21 Many teachers and classrooms that focused on special education students or English language learners could not participate in randomization because those students had to have a teacher with particular certification, and this most likely contributed to higher test scores in our randomization sample.

22 Actually, the covariance in teacher's students' baseline scores from one year to the next serves as an indicator of persistent sorting.

Table 3 also suggests that there was a considerable amount of persistent sorting (that is, some teachers consistently got the highest or lowest scoring students). To illustrate this, we estimated the variance in persistent differences in mean baseline test scores of students assigned to teachers.[23] Among the randomized teachers, the standard deviation of the persistent component of sorting (within-schools) was roughly .30 in both math and ELA. The persistent sorting seemed to be even higher for the non-randomized teachers in the MET project schools (.44 in math and ELA) and in the non-MET project schools (.380 in math and .397 in ELA).

Table 3

COMPARING RANDOMIZED MET PROJECT TEACHERS TO OTHER TEACHERS

	MET Project Schools		Non-MET Project Schools
Variable	Randomized Sample	Non-randomized Sample	
Teacher Value-Added:			
Mean value-added on state math tests	0.014	0.014	0.018
S.D. in value-added on state math tests	0.244	0.261	0.269
Signal S.D. in math effects	0.161	0.165	0.175
Mean value-added on state ELA tests	0.023	0.003	0.011
S.D. in value-added on state math tests	0.187	0.183	0.227
Signal S.D. in ELA effects	0.107	0.102	0.116
Classroom Mean Characteristics:			
Mean student baseline math scores	0.101	-0.040	-0.048
S.D. in student baseline math scores	0.486	0.590	0.652
Signal S.D. in baseline math sorting	0.410	0.509	0.552
Within-school S.D. in baseline math scores	0.382	0.520	0.492
Within-school signal S.D. in student sorting	0.297	0.437	0.380
Mean student baseline ELA scores	0.089	-0.048	-0.058
S.D. in student baseline ELA scores	0.475	0.598	0.661
Signal S.D. in baseline ELA sorting	0.392	0.504	0.549
Within-school S.D. in baseline ELA scores	0.387	0.535	0.513
Within-school signal S.D. in student sorting	0.300	0.440	0.397

(CONTINUED)

23 Specifically, we used the covariance in teachers' mean student baseline scores from 2008–09 and 2009–10 after adjusting for school fixed effects. This is analogous to our calculation of signal variance for value-added.

COMPARING RANDOMIZED MET PROJECT TEACHERS TO OTHER TEACHERS (CONTINUED)

Variable	MET Project Schools		Non-MET Project Schools
	Randomized Sample	Non-randomized Sample	
Teacher Characteristics:			
Mean teaching experience	8.214	7.510	7.624
S.D. in teaching experience	6.798	7.801	7.627
Proportion with higher degrees	0.361	0.406	0.202
Proportion with NBPTS certification	0.018	0.009	0.005
Proportion male	0.166	0.181	0.168
Proportion African American	0.345	0.392	0.232
Proportion Hispanic/Latino	0.060	0.068	0.126
Student Characteristics:			
Special education	9.2%	12.6%	17.3%
English language learner	19.9%	22.1%	42.7%
Hispanic	37.0%	37.9%	40.0%
African American	30.5%	38.2%	30.0%
White	32.5%	23.8%	30.0%
Male	50.1%	51.1%	51.0%
Special education missing	0.8%	0.5%	0.1%
Race missing	0.8%	0.7%	0.0%
Number of students	67,402	186,886	585,142
Number of teachers	1,181	3,802	17,153

Note: The value-added estimates and sorting on baseline scores were calculated only for those teachers with 10 or more students. In order to calculate the signal variance in value-added and sorting on baseline scores, we used the covariance at the teacher level between 2008–09 and 2009–10. The sample includes only teachers in grades 4–8.

The high level of sorting—persistent or not—is noteworthy for two reasons. First, although it obviously says nothing about the amount of sorting on unmeasured traits, it suggests that failing to control for observable traits, such as students' prior achievement, may lead to biased estimates of teacher effects (presuming that these traits are related to achievement gains). Second, there was a lot of sorting of students among the MET project teachers in the years prior to randomization. If the volunteers were limited to those who were already effectively randomly assigning students to teachers, then our test would not be generalizable to other classrooms where sorting was occurring. That does not seem to have been the case.[24]

Table 3 also reports the mean levels of experience, demographics, and credentials for the teachers in the three groups. Although they were similar in terms of gender and race/ethnicity, the MET project randomized teachers were somewhat more experienced (mean experience of 8.2 years as opposed to 7.5 and 7.6 for the other two groups), which may reflect the fact that MET project teachers had to have been teaching in the prior year. They had similar levels of higher degree completion as their counterparts in the MET project schools (36 versus 41 percent), although both groups had higher rates of higher degree completion than the teachers in the non-MET project schools (which was 20 percent).

Table 4 reports student characteristics for various subsamples of students. The first column reports characteristics of all the 4th–8th grade students in the randomized rosters who were assigned in the summer of 2010. In fact, we did not observe end-of-year test scores for 13 percent of these students. Therefore, the second column is our primary analysis sample, which is limited to those with end-of-year test scores on the state tests. The students with end-of-year 2011 scores had slightly higher achievement on the prior state tests, .139 versus .113 in math and .146 versus .122 in ELA. However, their demographic characteristics (race/ethnicity, gender) and program participation (special education and ELL status) were quite similar.[25]

In addition to the state tests, we collected other outcomes, such as student scores on supplemental assessments and students' self-reported academic persistence and enjoyment in class. Because these data were limited to those students who remained in the classroom of a MET project teacher, we lost data for 40 percent of the originally randomized sample. As reported in column 3, the subgroup of students with MET project outcomes had higher prior achievement in spring 2010—.200 versus .113 in math and .201 versus .122 in ELA—than the originally randomized sample. However, their demographics were again similar to the originally randomized sample.

In the fourth and fifth columns we report analogous results for the high school sample. However, for the high school students, we are lacking end-of-year state tests for students in many districts. As a result, we concentrated on the students with scores on the supplemental assessments administered by the study team. In high school, we lost slightly more than half of the original randomized sample (53 percent). As in 4th through 8th grades, the students who remained had slightly higher average achievement in the previous spring than the full sample that was originally randomly assigned: .104 versus .026 in math and .103 versus .029 in ELA.

24 Nevertheless, the sorting on the basis of observed student traits was somewhat higher among the non-randomized teachers in the MET project schools. Table 3 provides some hints as to why this might have occurred. While 9.2 percent of the randomized teachers' students were special education students in 2009–10, the percentages were higher for the non-randomized teachers in the MET project schools and the non-MET project schools: 12.6 percent and 17.3 percent, respectively. Moreover, the percentage of English language learners (ELLs) taught by randomized teachers was 19.9 percent, as opposed to 22.1 percent and 42.7 percent for the non-randomized teachers and teachers in non-MET project schools, respectively. To the extent that special education students and ELL students tend to be sorted into lower achieving classrooms, this may account for the greater degree of sorting outside the randomized MET project classrooms.

25 We did not have free and reduced-price lunch status for one of the districts and therefore did not include it in this table.

Table 4

EFFECTS OF ATTRITION ON MEAN STUDENT CHARACTERISTICS

	Grades 4–8			High School	
Variable	Randomized Sample	Randomized Sample with State Test Scores	Randomized Sample with MET Project Outcomes	Randomized Sample	Randomized Sample with MET Project Outcomes
Prior math score	0.113	0.139	0.200	0.026	0.104
Prior ELA score	0.122	0.146	0.201	0.029	0.103
Special education	0.086	0.080	0.076	0.078	0.073
English language learner	0.122	0.131	0.143	0.090	0.084
Hispanic	0.282	0.308	0.311	0.294	0.291
African American	0.298	0.309	0.295	0.295	0.311
White	0.232	0.257	0.284	0.225	0.307
Male	0.357	0.377	0.351	0.387	0.380
Special education missing	0.098	0.028	0.003	0.100	0.000
Race missing	0.050	0.000	0.003	0.035	0.000
Sample size	31,439	27,265	18,718	8,657	4,082

Note: The first three columns are reported for the main randomized sample, which includes students assigned to randomized teachers in grades 4 through 8. The last two columns are reported for the randomized high school sample, which includes all 9th grade students assigned to a randomized teacher. For the analysis of teacher impacts on students, an individual student could appear for more than one subject (e.g., math and ELA for elementary school students). To be consistent, this table also allows students to appear more than once. Each row of the table contains means of the variable indicated in the first column. Value of 0.000 for "race missing" in column 2 due to fact that no students were in that category. Values of 0.000 in fifth column due to rounding.

TREATMENT OF OUTLIERS

More than the observations and student surveys, the value-added data are prone to extreme values. Outliers can be useful—stretching a hypothesized relationship to its extremes—but only so long as they reflect legitimate differences in teacher performance. If they reflect factors other than the quality of instruction, they could lead us astray. After observing subsequent declines in achievement for students who appeared to have made large gains in a given teacher's classroom and after seeing unusual patterns of student responses, Jacob and Levitt (2003) concluded that as much as 2 percent of classrooms with the largest gains in student achievement in Chicago showed signs of cheating. Accordingly, for the remainder of the paper, we dropped teachers in the top 1 percent of the value-added distribution in math or ELA. In Appendix B, we report the sensitivity of our primary results to different amounts of trimming based on extreme values of value-added.

V. The Fidelity of the Experiment: Balance, Attrition, and Non-Sample Peers

In a randomized control trial (RCT) with a single treatment and a control group, it is conventional to test the degree to which the two groups differ in terms of observed characteristics at baseline—as a way to check whether the randomization procedures were followed and whether random assignment succeeded in creating treatment and control groups that are similar on pre-assignment variables. However, in our analysis, we are pooling across many different experiments, with a different "treatment" (that is, a different teacher) being assigned to each roster of students. Accordingly, we report the relationship between the assigned teacher's measure of effectiveness and the baseline characteristics of students. If randomization successfully equalized the student traits, then there should be no relationship between the assigned teacher's measured effectiveness and any observed student characteristic.[26]

To test for balance, we estimated the following regression for each of the student characteristics, X_i^l (for variables l=1 to L), using data at the student level:

$$(5)\ X_i^l = \pi_l \hat{\tau}^S_{jAssigned} + \phi_k^l + \varepsilon_i$$

where $\hat{\tau}^S_{jAssigned}$ is the measured effectiveness of the assigned teacher and ϕ_k^l represents the fixed effects for the randomization blocks.

The first column of **Table 5** reports the coefficient ($\hat{\pi}_l$) on the assigned teacher's measured effectiveness for each baseline characteristic for the sample of 4th through 8th graders. None of the coefficients on the individual characteristics is statistically distinguishable from zero at the .05 level, and the joint hypothesis test that all the coefficients were zero has a p-value of .163. In other words, we could not reject the hypothesis that assigned teachers' measured effectiveness was unrelated to all these characteristics (as we would have expected if teachers had been randomly assigned).

The second column of Table 5 reports a similar set of relationships for the subsample of students for whom we have student achievement measures in spring 2011. Even if there were no relationship at the moment of random assignment, such a relationship could have re-emerged within the subset of students for whom we were able to track down end of year scores. Since this is our sample for most of the following analyses, we tested for balance in this subgroup. Again, there was no relationship between the assigned teacher's measured effectiveness and any of the individual student characteristics. The joint hypothesis test assuming that all the coefficients were zero had a p-value of .282. The third column is limited to the subset of students who had data on one of the outcomes collected specifically for the MET project—supplemental assessments and student survey results. Again, none of the individual coefficients were statistically significant. Moreover, the joint hypothesis test could not reject the absence of a relationship between the assigned teacher's measured effectiveness and any of the student characteristics (p-value of .106).

26 Except, of course, in the case of an unlucky draw. Even if the assignment process truly was random, some of the differences in baseline characteristics could exceed the threshold of statistical significance.

Table 5

BALANCE: ASSIGNED TEACHER EFFECTIVENESS AND STUDENT CHARACTERISTICS

	Grades 4–8			High School	
Variable	Randomized Sample	Randomized Sample with State Test Scores	Randomized Sample with MET Project Outcomes	Randomized Sample	Randomized Sample with MET Project Outcomes
Prior math score	0.238 (0.185)	0.317 (0.210)	0.304 (0.231)	0.033 (0.167)	-0.091 (0.213)
Prior ELA score	0.276* (0.166)	0.359* (0.187)	0.171 (0.190)	0.039 (0.193)	-0.005 (0.230)
Special education	-0.021 (0.057)	-0.049 (0.061)	0.045 (0.052)	-0.045 (0.051)	0.001 (0.066)
English language learner	-0.057 (0.059)	-0.070 (0.066)	-0.100 (0.094)	0.107 (0.081)	0.0910 (0.083)
Hispanic	-0.032 (0.047)	-0.057 (0.052)	-0.083 (0.065)	-0.001 (0.056)	0.000 (0.100)
African American	0.043 (0.053)	-0.005 (0.044)	0.014 (0.063)	0.071 (0.086)	0.161 (0.136)
White	-0.007 (0.039)	-0.009 (0.044)	-0.037 (0.057)	-0.087 (0.082)	-0.145 (0.154)
Male	-0.087 (0.054)	-0.108* (0.065)	-0.104 (0.083)	0.021 (0.091)	-0.043 (0.158)
Special education missing	-0.083* (0.049)	-0.024 (0.026)	0.013* (0.007)	0.027 (0.066)	-0.002 (0.003)
Race missing	-0.047 (0.035)	NA	0.013* (0.007)	-0.029 (0.028)	-0.002 (0.003)
p-value on joint H_o that all are zero	0.163	0.282	0.106	0.197	0.695

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The first three columns are reported for the main randomized sample, which includes students assigned to randomized teachers in grades 4 through 8. The last two columns are reported for the randomized high school sample, which includes all 9th grade students assigned to a randomized teacher. The table contains coefficients on assigned teacher effectiveness with different student characteristics as the dependent variable. All specifications also include fixed effects for randomization block. The joint hypothesis test was estimated using the "seemingly unrelated regressions" or SURE model proposed by Zellner (1962). Assigned teacher effectiveness is the prediction of the assigned teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. Standard errors were calculated allowing for clustering within randomization block. "NA" for "race missing" due to the fact that no students were in that category.

The last two columns refer to the sample of high school students. There was no relationship between the assigned teacher's measured effectiveness and any of the student characteristics at the moment of randomization for the high school sample. Moreover, the p-value on the joint hypothesis that all the coefficients were zero was .197. The last column reports the same relationships for the high school sample with end-of-course tests in math, ELA, or biology, when the project administered the QualityCore assessments from ACT. Again, there was no evidence of a relationship between the assigned teacher's measured effectiveness and student characteristics at baseline (p-value of .695).

ATTRITION

The first column of **Table 6** reports the proportion of the initially randomized sample with various outcomes: state test scores and MET project outcomes. Eighty-seven percent of the 4th through 8th grade students on the initial randomization lists had state test outcomes at the end of the year. Because we were only able to collect the MET project outcomes for those who remained in a MET project classroom, those data were available for a smaller percentage of students, 60 percent for the 4th through 8th grade sample and 47 percent for the 9th grade sample.

The second column of Table 6 reports the relationship between each indicator of data availability at the student level and the students' assigned teacher's measured effectiveness. There was no statistically significant relationship between the assigned teacher's measured effectiveness and the availability of the subsequent state achievement measures.

Table 6

ATTRITION AND ASSIGNED TEACHER EFFECTIVENESS

	Proportion	Coefficient on Assigned Teacher Effectiveness
Grades 4–8 Randomized Sample:		
Student has state test scores	0.866	-0.012 (0.033)
Student has MET project outcomes	0.595	-0.070 (0.143)
High School Randomized Sample:		
HS student has MET project outcomes	0.472	-0.077 (0.144)

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The first two rows are reported for the main randomized sample, which includes students assigned to randomized teachers in grades 4 through 8. The third row is reported for the randomized high school sample, which includes all 9th grade students assigned to a randomized teacher. The table contains coefficients on assigned teacher effectiveness with dummy variables indicating if the student had outcome data as the dependent variable. All specifications also included fixed effects for randomization block. Assigned teacher effectiveness is the prediction of the assigned teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. Standard errors were calculated allowing for clustering within randomization block.

NON-SAMPLE PEERS

In health trials, there is no need to measure the characteristics of patients outside the study sample. For instance, a cancer patient's reaction to a new medication is not influenced by the characteristics of others taking the same drug, especially if they were not part of the clinical trial. However, in many educational interventions, this may not be the case, since other students in a classroom—even if they were not part of the random assignment—could influence the performance of members of the study sample. Even with random assignment, peer effects could re-introduce bias if those assigned to more effective teachers ended up enjoying more or less positive peer influences.

With random assignment, assigned teacher effectiveness should not be related to any traits of the randomized students, measured or unmeasured. However, given the entry of non-sample members into MET project classrooms and the exit of sample members into other classrooms, random assignment does not guarantee that assigned teacher effectiveness is unrelated to peer characteristics. We can use our data to test if there is any relationship of assigned teacher effectiveness to measured peer characteristics. Whereas Table 5 focused on the relationship between an assigned teacher's measured effectiveness and baseline characteristics of each student randomized on a randomized roster, **Table 7** looks at the relationship between an assigned teacher's measured effectiveness and the mean characteristics of the actual classroom peers for each sample member in spring 2011. Looking at each of the individual characteristics, there is no evidence that assigned teacher effectiveness was related to the observable characteristics of classroom peers. And the joint hypothesis test would not lead us to reject the hypothesis that all the coefficients are equal to zero (p-value of .405).

The bottom panel of Table 7 reports similar estimates for the high school sample. Again, assigned teacher effectiveness is only related to one of the peer measures (absences) and at a marginal level of statistical significance. The joint test finds that we cannot reject the joint hypothesis that all the peer measures are unrelated to assigned teacher effectiveness (p-value of .321).

In sum, despite the movement into and out of the randomized classrooms, the peers to whom randomized students were exposed were not systematically different based on the effectiveness of their randomly assigned teacher.

Table 7

RELATIONSHIP BETWEEN ASSIGNED TEACHER EFFECTIVENESS AND SUBSEQUENT PEER CHARACTERISTICS

	Baseline Mean	Coefficient of Assigned Teacher Effectiveness
Grades 4–8 Randomized Sample:		
Baseline math scores of actual peers	0.105	0.246 (0.164)
Baseline ELA scores of actual peers	0.090	0.245 (0.152)
% special education of actual peers	0.083	-0.038 (0.042)
% ELL of actual peers	0.118	-0.050 (0.052)
% Black of actual peers	0.289	0.059 (0.048)
% Hispanic of actual peers	0.269	-0.051 (0.035)
Number of student absences	7.578	-0.153 (1.820)
Number of suspensions	0.061	-0.022 (0.026)
p-value on joint H_o that all are zero		0.405
High School Randomized Sample:		
Baseline math scores of actual peers	0.007	0.014 (0.164)
Baseline ELA scores of actual peers	0.011	-0.125 (0.194)
% special education of actual peers	0.076	-0.002 (0.039)
% ELL of actual peers	0.090	0.081 (0.062)
% Black of actual peers	0.272	-0.017 (0.070)
% Hispanic of actual peers	0.254	-0.069 (0.046)
Number of student absences	11.17	-1.289 (1.518)
Number of suspensions	0.151	-0.020 (0.050)
p-value on joint H_o that all are zero		0.321

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: Although teachers were randomly assigned to classroom rosters, subsequent movements of students may have changed the peers available to students. The table contains coefficients on assigned teacher effectiveness with measures of average classroom peer characteristics as the dependent variable. All specifications also included fixed effects for randomization block. The joint hypothesis test was estimated using the "seemingly unrelated regressions" or SURE model proposed by Zellner (1962). Assigned teacher effectiveness is the prediction of the assigned teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. Standard errors were calculated allowing for clustering within randomization block.

VI. The Impact of Assigned and Actual Teacher Effectiveness

Table 8 reports the effect of the *assigned* teacher's measured effectiveness on a student's actual teacher's measured effectiveness—the first stage model of the instrumental variables estimator. The first row is reported for the 4th through 8th grade sample, and it pools the results for math and ELA teachers. If there were perfect compliance, 100 percent of assigned teacher effectiveness would have translated into actual teacher effectiveness and the coefficient would have been one. Instead, the coefficient estimate was .293, implying that a one-unit improvement in the assigned teacher's measured effectiveness led to a statistically significant .293 unit increase in the actual teacher's measured effectiveness.[27] Despite the movement of students and teachers afterward, the random assignment process did lead to differences in the effectiveness of students' actual teachers, which we can use to infer the impact of a teacher's measured effectiveness on outcomes. Yet the full impact of the randomly assigned teachers' effectiveness was diluted by non-compliance. The second and third rows of Table 8 report the results separately for ELA and math. A slightly larger share of the assigned teacher effectiveness filtered through to actual teacher assignments in ELA than in math: .326 versus .281.

The next six rows of Table 8 report the first stage coefficients separately for each school district. Due to higher levels of compliance with the randomly assigned rosters, the effectiveness of the assigned teachers were more likely to filter through to actual teacher effectiveness in Dallas and Charlotte-Mecklenburg (.744 and .517, respectively) than in Hillsborough, Memphis, and New York (.380, .277, and .157, respectively). There was no statistically significant relationship between assigned teacher effectiveness and actual teacher effectiveness in Denver because of the low compliance rate in that district.[28] Fortunately, the Denver sample represented only 6 percent of the randomly assigned students in our sample.

TEACHER IMPACTS ON STATE TEST SCORES

Table 9 reports the instrumental variable estimates of teacher effectiveness on student achievement. As described above, these estimates combine the two effects of measured effectiveness of the randomly assigned teacher—on the effectiveness of the teachers actually working with students and on student achievement—to infer the impact of a student's actual teacher's measured effectiveness. The first three columns contain results from our preferred specification, while including student background characteristics as covariates. To maximize statistical power, we first pooled the math and ELA results.[29] When math and ELA results are combined, the coefficient on actual teacher effectiveness on student achievement was .955, with a standard error of .123.

27 As noted in the text, if every student's actual teacher was his or her assigned teacher, i.e., perfect compliance, the coefficient would be one. If schools did not comply at all with the experiment and ignored teacher assignment, the coefficient would be zero.

28 In many of Denver's schools, one or two of the teachers in a particular grade or subject had certification necessary for instructing Spanish speaking students and the remaining teachers did not. Although teachers and principals were informed at the beginning of the project that participation would require randomly assigning rosters among teachers in randomization blocks, some randomization blocks included the teachers with and without the Spanish language certification. These teachers could not comply with the randomization—since doing so could mean that the Spanish-speaking students did not have a Spanish-speaking teacher.

29 In the pooled specification, we used randomization block by subject fixed effects and allowed the coefficients on baseline scores in math and ELA to vary by the subject of the dependent variable score.

Table 8

RELATIONSHIP BETWEEN ASSIGNED AND ACTUAL TEACHER EFFECTIVENESS (FIRST STAGE)

Grade 4–8 Randomized Sample by Subject:	
Math and ELA (stacked)	0.293*** (0.045)
Math	0.281*** (0.051)
ELA	0.326*** (0.058)
Grade 4–8 Randomized Sample by District (Math & ELA stacked):	
Dallas	0.744*** (0.048)
Charlotte-Mecklenburg	0.517*** (0.065)
Hillsborough	0.380*** (0.072)
Memphis	0.277*** (0.077)
New York	0.157* (0.086)
Denver	-0.079 (0.214)
High School Randomized Sample by Subject:	
Math, ELA, and Biology (stacked)	0.447*** (0.114)
Math	0.536*** (0.139)
ELA	0.581*** (0.104)
Biology	0.354* (0.183)

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The first two panels of estimates are reported for the main randomized sample, which includes students assigned to randomized teachers in grades 4 through 8. The third panel of estiamtes is reported for the randomized high school sample, which includes all 9th grade students assigned to a randomized teacher. The dependent variable is each student's actual teacher effectiveness. The table reports the coefficient from a regression of actual teacher effectiveness on assigned teacher effectiveness. Assigned teacher effectiveness is the prediction of the assigned teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

With a t-statistic value of 8, we can reject the hypothesis that the coefficient is zero. In other words, the results are inconsistent with the idea that the measured effectiveness of the assigned teachers has no causal effect on student achievement.

In fact, we cannot reject the hypothesis that the coefficient is equal to one. The 95 percent confidence interval extends from .714 to 1.196. Because one is included in the confidence interval, we cannot reject the hypothesis that a one-unit change in teachers' measured effectiveness corresponds to an average change of one unit in the true effects of the teachers when comparing teachers teaching the same subject and grade in the same school.

When estimated separately for math and ELA, the coefficients are 1.039 and .697, respectively. In both cases, we could reject the hypothesis that the coefficient on measured teacher effectiveness was equal to zero. And in neither case could we reject the hypothesis that the coefficient was equal to one. Moreover, we cannot reject that the two are equal (p-value of .18). However, the ELA coefficient is both smaller and less precisely estimated than the math coefficient, with a 95 percent confidence interval on the ELA coefficient ranging from .280 to 1.114. Thus, while we have no strong evidence that ELA differs from math, our results for ELA are much less precise and provide limited evidence about the meaning of a unit change on the effectiveness scale for ELA teachers.

The next three columns report similar results, but they exclude any controls for student baseline scores during the randomization year. Since students were randomly assigned to teachers, this specification also provides unbiased estimates of the causal effect of teachers, albeit the estimates are less precise since the inclusion of covariates explains some of the variance in the outcome. For all three, we could reject the hypothesis of no impact, while failing to reject the hypothesis that the effectiveness measures have a one-unit impact on student achievement.

The last three columns include as covariates the actual peer characteristics of students following random assignment. Because the actual peers included students who were not part of the original random assignment, we cannot rely on the random assignment to ensure unbiased estimates of the effects of peers. Nevertheless, the inclusion of measured characteristics of actual peer controls has little impact on the estimates. This is consistent with the earlier findings in Table 7 that assigned teacher effectiveness was unrelated to actual peer characteristics to ensure unbiased estimates of the effects of peers.

VARYING THE INSTRUMENT SET

As a robustness check, we used a range of different instrumental variables for predicting a teacher's actual teacher effectiveness. The instruments used in the prior table included interactions between assigned teachers' measured effectiveness and a binary indicator for each school to account for differing compliance by school. In the first and second columns of **Table 10**, we report the result of using assigned teacher effectiveness alone (with no interactions) and then teacher effectiveness by district interactions—with no interactions by school. The point estimates, .884 and 1.040, are similar to that observed with the school interactions, although the standard errors are twice as large. The third column allows for interactions at the randomization block level—therefore, including even more interactions than used in **Table 9**. The results—a coefficient of 1.020 and a standard error of 0.117—are similar to our baseline specification (reproduced in the fourth column of Table 9), suggesting that there was little to be gained from allowing for within-school variation in compliance—that is, most of the variation in compliance was at the school level, as one might expect.

Table 9

INSTRUMENTAL VARIABLE ESTIMATES OF TEACHER EFFECTS ON STUDENT ACHIEVEMENT

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math and ELA (Stacked)	Math	ELA	Math and ELA (Stacked)	Math	ELA	Math and ELA (Stacked)	Math	ELA
Expected student achievement based on teacher's effectiveness	0.955*** (0.123)	1.039*** (0.138)	0.697*** (0.213)	1.342*** (0.262)	1.566*** (0.313)	0.919** (0.368)	0.935*** (0.116)	1.013*** (0.126)	0.657*** (0.210)
Controls for student's prior achievement and demographics?	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Controls for actual peer characteristics?	No	No	No	No	No	No	Yes	Yes	Yes
Observations	27,255	13,457	13,798	27,255	13,457	13,798	27,255	13,457	13,798
R-squared	0.684	0.705	0.666	0.277	0.293	0.260	0.686	0.708	0.668
Number of randomization blocks	619	309	310	619	309	310	619	309	310

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school level indicators as instruments. Columns (1), (4), and (7) treat student achievement in math and ELA as separate observations, with separate randomization blocks. Prior achievement measures are student scores in math and ELA on state tests in the prior year, which are interacted with the dependent variable subject in columns (1), (4), and (7). All specifications also included fixed effects for randomization block. Standard errors were calculated allowing for clustering within teacher.

Instrumental variable estimators can be biased when there are a large number of weakly correlated instruments, particularly when the F-statistic testing the joint significance of the instruments from the first stage estimation is small in magnitude (Staiger and Stock, 1997). The minimum F-statistic in our analyses across our base specification with interactions between teacher effectiveness raters and school dummies and our alternative specifications was 38, implying that weak instrument bias is not a concern.

Table 10

INSTRUMENTAL VARIABLE ESTIMATES OF TEACHER EFFECTIVENESS ON STUDENT ACHIEVEMENT: VARYING THE INSTRUMENTAL VARIABLE SETS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent Variable: Math and ELA Stacked						
Expected student achievement in teacher's class based on teacher's effectiveness in prior year	0.884*** (0.259)	1.040*** (0.183)	1.020*** (0.117)	0.955*** (0.123)	1.148*** (0.149)	0.807*** (0.293)	0.940*** (0.311)
Interactions with assigned teacher effectiveness included in instrument set	None	District	Randomization Block	School	School	School	School
Measures used in creating teacher effectiveness	VA, Observation, Stud. Survey	VA, Observation, Stud. Survey	VA, Observation, Stud. Survey	VA, Observation, Stud. Survey	VA Only	Observation Only	Survey Only
Observations	27,255	27,255	27,255	27,255	27,255	27,255	24,415
R-squared	0.684	0.684	0.684	0.684	0.684	0.681	0.685
Number of randomization blocks	619	619	619	619	619	619	619

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Student achievement in math and ELA are treated as separate observations, with separate randomization blocks. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on a combination (as noted in the table) of value-added, student surveys, and observations in the prior school year. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers and interactions (as noted in the table) as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

VARYING THE COMPONENTS OF THE EFFECTIVENESS RATING

While we used the composite measure of effectiveness above, the last three columns of Table 10 use each of these components individually.[30] The results are qualitatively similar to those we saw with the full composite: the estimates of γ_1 are large and not significantly different from one. However, because classroom observations and student perception surveys are much less strongly predictive of student achievement gains, the standard errors on these are considerably higher (roughly of .3 rather than .149 for the value-added measure).

30 For each component we re-estimated a modified version of Model (3), which excluded the other two components to generate a teacher's measured effectiveness based on a single component. For example, for a measure of effectiveness using only FFT, we estimated the model

$$(3b)\quad \hat{\tau}^{S}_{j2008-09} = \beta_o + \beta_2 FFT^{adj}_{j2009-10} + \beta_4 Z_j + \eta_{jt}$$

The estimates of achievement gain only and Tripod only ratings were calculated similarly. When used on their own, value-added, FFT, and Tripod each predicted narrower differences between teachers than the composite measure did.

BY GRADE LEVEL

In **Table 11**, we report the results separately for elementary (grades 4 and 5) and middle school grades (grades 6 through 8). The point estimates are similar—.994 and .892, respectively. Moreover, we could not reject the hypothesis that the coefficients are the same for middle and elementary school grades (p-value of .693).

Table 11

INSTRUMENTAL VARIABLE ESTIMATES OF TEACHER EFFECTIVENESS ON STUDENT ACHIEVEMENT: BY GRADE LEVEL

Coefficient on expected student achievement in teacher's class:	Math and ELA Stacked
Grades 4 and 5	0.994*** (0.153)
Grades 6 through 8	0.892*** (0.209)
p-value for test of equal coefficients	0.693
Observations	27,255
R-squared	0.684
Number of randomization blocks	619

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school level indicators as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

TESTING FOR BIAS IN THE ABSENCE OF STUDENT CONTROLS AND ADJUSTMENTS FOR VOLATILITY

The evidence in Tables 9 through 11 implies that the adjustments for student baseline characteristics and year-to-year volatility generated measures of effectiveness that were, on average, not statistically different from what we observed following random assignment. But would the results have been different if we had not adjusted for student baseline characteristics or if we had not adjusted for volatility in test scores?

The first column of **Table 12** repeats the results using our preferred value-added model from Table 9. The second column uses teacher effect estimates not adjusted for students' prior achievement or characteristics. That is, instead of using value-added to create the measure of effectiveness, we used teacher end-of-year 2009–10 mean student test scores (and unadjusted student surveys and observation scores) to predict a teacher's mean

end-of-year scores from 2008 to 2009. Given the apparent sorting on student baseline scores reported in Table 1, we would expect the estimates from such a model to be biased. Indeed, this appears to be the case. The coefficient (.228) is statistically different from zero and from one. The confidence interval covers a range of values that are all far less than one, from .146 to .310. In other words, if we used end-of-year scores to assess teachers' effectiveness and failed to adjust for students' prior achievement, then we would be overstating the differences between teachers.[31]

The third column uses the raw value-added estimates emerging from equation (1), $\hat{\tau}_j^S$, without adjusting for volatility using equation (4).[32] With a point estimate of .430 and a standard error of .057, we could again reject the hypothesis that the coefficient is equal to one. In other words, by using the raw value-added estimates and not using the shrinkage factors estimated in equation 4, we see differences in achievement following random assignment about half as large as we would have predicted.

This has practical significance, since many districts and states do not adjust the value-added estimates for volatility. If a teacher has a raw value-added estimate 5 points higher than another teacher in his or her school teaching the same grade and subject, then the expected difference in achievement if those two teachers were to work with another group of students would be about half as large. However, since the shrinkage factor we use is a constant for each outcome, it would not change the relative ranking of teachers for a given outcome. Rather, failing to account for shrinkage is essentially a problem of scaling, being overoptimistic about the magnitude of the difference in achievement one would expect to see between two teachers in the future. It is when a school system pools value-added estimates for teachers measured with different tests, with differing degrees of volatility (e.g., math teachers and ELA teachers) that such bias could change not only the scaling but the ranking of teachers as well.[33]

TESTING ALTERNATIVE VALUE-ADDED MODELS

In the past, policymakers have had to choose a list of student characteristics for which to adjust without knowing the magnitude of the trade-offs involved. For example, a student's poverty status may signal a shortage of vital supports at home, such as adequate study space or parental help with homework. Many argue that unless we control for students' poverty status, we will unfairly blame teachers for student underperformance driven by factors beyond their control. On the other hand, people have argued that controlling for student poverty status can signal different expectations for different students. Florida has gone so far as to create regulations forbidding the use of poverty status and race in value-added models for teacher effects.

31 It is worth nothing that even this biased estimator—which fails to control for any student characteristics and simply uses end-of-year scores to evaluate teacher performance—does provide predictive power with respect to student achievement following random assignment. The differences are simply overstated.

32 The sample size in column (3) is reduced, since we lost all the sample members who had no value-added scores in the prior years. Nearly all of the assigned teachers had value-added scores, but some of the actual teachers did not. When we were using equation (3) to generate the composite measure adjusted for volatility, we could generate a prediction even in the absence of a value-added estimate. However, if we are not using equation (3), there is no way to impute a predicted effectiveness for those teachers without value-added.

33 The bias resulting from failing to control for prior achievement is different. It would not only lead one to overstate the difference. It would also have an impact on the relative rankings of teachers, given sorting on the basis of prior achievement.

Table 12

TESTING ALTERNATIVE METHODS OF ESTIMATING TEACHER IMPACT ON STUDENT ACHIEVEMENT

	(1)	(2)	(3)
	Math and ELA Stacked	**Math and ELA Stacked**	**Math and ELA Stacked**
Expected student achievement in teacher's class based on teacher's effectiveness in prior year	0.955*** (0.123)	0.228*** (0.042)	0.430*** (0.057)
Method for estimating teacher impact on student acheivement	Regression composite, control for peer effects	Regression composite, mean end-of-year student score, observation, and student survey **with no controls**	Value-added with no shrinkage, includes control for peer effects
Observations	27,255	27,255	24,488
R-squared	0.684	0.682	0.688
Number of randomization blocks	619	619	613

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, using the method noted in the table. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school level indicators as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

Because of the second-year random assignment, we have a unique opportunity to ask: Which value-added models best predict student achievement following random assignment? Whenever we added an additional set of control variables to our value-added model, a portion of the inferred variance in teacher effects was "taken away" and attributed to the control variables, not to teachers.[34] But that may not always be appropriate, since some of that variance could be legitimately attributable to teachers.

That suggests a natural test: Suppose that we start with our primary model (which includes all three sets of controls—student baseline scores, demographics, and peer effects) and then sequentially add back that portion of the teacher effect that was subtracted off with each new set of control variables. We could test which, if

34 In a linear regression, only that portion of the teacher effect that is not a linear function of the control variables (i.e., that which is orthogonal to the teacher effect) is used to infer the effect of teachers. Our value-added models did not include teacher fixed effects but rather used the two-step process of controlling for variables in the first step and then averaging adjusted score (residuals) to obtain value-added estimates. With fixed effects, we could avoid removing the control variables from the teacher effects for student level variables but with a single year of data we cannot used fixed effects with teacher-level variables. Many states and districts use a similar approach to the one used in our analyses; however, some do use fixed effects. In models that control for only student-level variables, models with and without teacher fixed effects yield nearly identical results (Ehlert et al., under review), so results on student-level demographics for such models are likely to be similar to those reported in Table 13.

any, of those incremental parts would have been predictive of student achievement following random assignment. If they contain a portion of the true teacher effect, they should be able to predict outcomes following random assignment. If not, they should be unrelated.

Table 13 reports the results of that analysis. We have four different measures of effectiveness for each teacher, based on different models. Each is a composite of our three measures—achievement gains, classroom observations, and student surveys. However, we adjusted each of the three measures for different sets of covariates:

$\hat{\bar{\tau}}_j^0$ Our primary model, which included controls for student baseline scores, students' baseline demographics, and mean peer characteristics.

$\hat{\bar{\tau}}_j^1$ A second model, which controls for student baseline scores, students' baseline demographics, but not mean peer characteristics.

$\hat{\bar{\tau}}_j^2$ A third model, which controls for student baseline scores but not students' baseline demographics or mean peer characteristics.

$\hat{\bar{\tau}}_j^3$ A fourth model, which does not control for student baseline scores, students' baseline demographics, or mean peer characteristics.

The first column uses our primary model, $\hat{\bar{\tau}}_j^0$, to predict teacher effects. It is possible that we "overcontrolled" in this model, and the component that was removed by peer effects could also predict additional differences in teacher performance under randomization.[35]

Therefore, the second column includes $\hat{\bar{\tau}}_j^0$ as well as $\hat{\bar{\tau}}_j^1 - \hat{\bar{\tau}}_j^0$, the portion of the teacher effect that was subtracted off when controlling for peer effects. We estimate the coefficient on this additional component in the same way as we have in other specifications, using the additional component for the assigned teacher as an instrument for the additional component of the student's actual teacher. The coefficient on this component is also near one and significantly different from zero. In other words, the component removed by controls for peer effects does seem to reflect causal teacher effects—not just factors outside a teacher's control. It is a significant predictor of teacher impacts on student test scores following randomization. This suggests that we may have been overcontrolling when we added controls for peer effects, capturing a portion of the effect attributable to teachers.

Column (3) repeats the exercise for the component removed by student demographics. It includes three variables: $\hat{\bar{\tau}}_j^0$, $\hat{\bar{\tau}}_j^1 - \hat{\bar{\tau}}_j^0$, $\hat{\bar{\tau}}_j^2 - \hat{\bar{\tau}}_j^1$. Although the coefficients on $\hat{\bar{\tau}}_j^0$ and $\hat{\bar{\tau}}_j^1 - \hat{\bar{\tau}}_j^0$ are both statistically different from zero and not statistically different from one, the component associated with student demographics, $\hat{\bar{\tau}}_j^2 - \hat{\bar{\tau}}_j^1$, is not statistically different from zero. In other words, the component removed by controls for student demographics does not seem to contain a portion that is associated with causal teacher effects—it is not

35 It is somewhat counterintuitive that our primary measure could accurately predict a one-unit difference in student outcomes yet still be overcontrolling for peer effects. Here's the intuition. One can decompose a teacher's true causal effectiveness into two components: One component that is correlated with classroom peer characteristics (due to sorting of effective teachers to classrooms with higher prior achievement scores); and the remaining component that is independent of peer characteristics. Our primary measure of effectiveness removes the component that is related to peer characteristics, but the remaining component is still correct on average because it is not related to peer characteristics by construction—some teachers (teaching classrooms with higher prior achievement scores) will have causal effects greater than our primary measure, while others (teaching classrooms with lower prior achievement scores) will have casual effects less than our primary measure. If we had not overcontrolled, we could have identified these additional causal differences between teachers that are related to peer characteristics.

Table 13

TESTING FOR BIAS IN COMPONENTS REMOVED BY VALUE-ADDED MODELS WITH INCREASINGLY RICH CONTROLS

	(1)	(2)	(3)	(4)
Variable	Math and ELA Stacked	Math and ELA Stacked	Math and ELA Stacked	Math and ELA Stacked
Expected student achievement in teacher's class based on teacher's effectiveness in prior year, controlling for student baseline scores, demographics, and peer effects	0.955*** (0.123)	0.947*** (0.115)	0.934*** (0.127)	0.879*** (0.131)
Component removed by controls for peer effects		1.150*** (0.336)	0.932*** (0.337)	0.636 (0.424)
Additional component removed by controls for student demographics			-0.025 (0.521)	0.015 (0.508)
Additional component removed by controls for student baseline scores				0.047 (0.042)
Observations	27,255	27,255	27,255	27,255
R-squared	0.684	0.684	0.684	0.685
Number of randomization blocks	619	619	619	619

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. See text for description of how additional components were calculated. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness and each of the components, using the effectiveness and components of randomly assigned teachers interacted with school-level indicators as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

a significant predictor of teacher impacts on student test scores following randomization. However, because the standard error is quite large and the confidence interval includes one, we also cannot reject the view that this component contains a sizable portion attributable to teacher effectiveness.[36] Unfortunately, our evidence cannot resolve the long debate over the inclusion of student demographics.

Column (4) adds the component associated with student baseline scores, $\hat{\tilde{\tau}}_j^3 - \hat{\tilde{\tau}}_j^2$. As with student demographics, the component associated with the baseline scores of individual students was not significantly related to the performance of students following random assignment. Unlike with student demographics, however, the effect of this component could be estimated fairly accurately (with a standard error of .04). This provides strong evidence that controls for student baseline scores are removing variation that is not attributable to teacher effectiveness.

In sum, we are comfortable recommending that school systems should control for student baseline scores, since this component is not predictive of outcomes following random assignment. At least for making comparisons within schools, the evidence also recommends against controlling for mean characteristics of students, since this component does seems to be predictive of outcomes following random assignment. However, the peer effect controls may be important to eliminate bias in between-school comparisons—a possibility we could not test. As a result, we cannot make a recommendation on the use of peer effects without further evidence. Finally, the case for or against controlling for demographic characteristics of students remains ambiguous.

IMPACTS ON OTHER STUDENT OUTCOMES

In **Table 14**, we report estimates of the impact of effective teachers on student outcomes other than state test scores. Like the state test scores, all of these measures have been standardized to have mean zero and standard deviation of one at the student level. There is no reason to expect the coefficient to be equal to one. However, a coefficient that is statistically different from zero implies that teachers who are effective at raising state test scores also have a causal impact on other student outcomes.

The sample size is smaller and attrition was higher for these outcomes. Nevertheless, teachers who were more effective on state tests also seemed to raise student achievement on the supplemental tests by .661 standard deviations (about two-thirds of a standard deviation). The item format and domains of the tests were distinct from the state tests. For example, the supplemental test in math, the Balanced Assessment in Mathematics, probed for students' conceptual understanding of math topics with very different item formats than used on state tests. The supplemental test in ELA required students to write short-answer responses after reading short passages, rather than simply answer multiple choice questions.

When the project conducted the student survey during the second year, we added some questions to explore the impacts of effective teachers on student outcomes other than those that could be measured on a test. As summarized in Tough (2012), there has been a burgeoning interest in research and policy circles in so-called non-cognitive outcomes, reflected in student motivation and persistence. The remaining columns of Table 14 report impacts on Angela Duckworth's measure of academic grit (which is intended to measure students'

36 The large standard error reflects the fact that relatively little of the variance in estimated teacher effects is attributable to demographic components. Whenever student baseline scores are included in the model, then the correlation between teacher effect measures with and without student demographics was above .95.

Table 14

INSTRUMENTAL VARIABLE ESTIMATES OF TEACHER EFFECTIVENESS ON OTHER OUTCOMES (RESTRICTED SAMPLE)

	(1)	(2)	(3)	(4)	(5)
Math and ELA stacked:	**Supplemental Test**	**Grit**	**Implicit Theory of Intelligence**	**Student Effort**	**Student Enjoys Class**
Expected student achievement in teacher's class based on teacher's effectiveness in prior year	0.661*** (0.153)	-0.029 (0.216)	0.331 (0.220)	0.199 (0.214)	0.966*** (0.285)
Observations	16,642	17,476	17,234	17,706	17,473
R-squared	0.490	0.065	0.182	0.091	0.129
Number of randomization blocks	553	559	558	559	559

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is the student outcome indicated at the top of each column. These outcomes are available only for a restricted sample of students, those who enrolled in a MET project teacher's classroom. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year.The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school-level indicators as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

willingness to persist on cognitively challenging problems), Carol Dweck's implicit theory of intelligence measure (which asks children to report the degree to which they believe success is a result of effort rather than fixed ability), students' self-reported level of effort in class, and students' level of enjoyment in class. Of these, the only statistically significant impact was on student enjoyment in class, where a teacher predicted to raise student achievement on state tests by one standard deviation raised students' enjoyment in class by .966 standard deviations. Students randomly assigned to more effective teachers reported enjoying class more.

RESULTS FOR HIGH SCHOOL

Because many of the MET project districts did not use end-of-course tests in high school, we were limited to the MET project outcomes for 9th grade students. As a result, the attrition rate was much higher at the high school level. Although we found no relationship between baseline student characteristics and assigned teacher effectiveness among those with MET project outcomes, we remain cautious in attaching too much weight to the high school results.

Table 15

INSTRUMENTAL VARIABLE ESTIMATES OF TEACHER EFFECTIVENESS IN HIGH SCHOOL

	(1)	(2)	(3)	(4)	(5)
	Math, ELA, and Biology Stacked	Math, ELA, and Biology Stacked	Math	ELA	Biology
Expected student achievement in teacher's class based on teacher's effectiveness in prior year	0.828*** (0.292)	0.629** (0.281)	0.578 (0.658)	1.049** (0.512)	0.961** (0.402)
Controls for student's prior achievement and demographics?	Yes	Yes	Yes	Yes	Yes
Controls for actual peer characteristics?	No	Yes	No	No	No
Observations	3,877	3,877	1,020	1,596	1,261
R-squared	0.381	0.381	0.294	0.384	0.423
Number of randomization blocks	149	149	49	55	45

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Note: The sample consists of all randomized students in grade 9. The dependent variable is student achievement on ACT's QualityCore assessment following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added on the QualityCore assessments from the 2010–11 school year, based on value-added on the QualityCore, student surveys, and observations in the prior school year. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school-level indicators as instruments. All specifications also included fixed effects for randomization block and controls for students' prior achievement on state tests and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

We report the high school results in Table 14. In the first column, we combine the results across three subjects: math, ELA, and biology. The instrumental variables estimate of the effect of actual teacher effectiveness was .828, with a standard error of .29. Like the results for earlier grades, we can reject the hypothesis of no impact and cannot reject the hypothesis that a one-unit increase in estimated teacher effectiveness has a causal effect of one unit on student achievement. In the second column, we include controls for actual peer characteristics. Adding the controls for actual peers lowers the coefficient somewhat to .629, with a standard error of .281.

In the last three columns of **Table 15**, we report the results separately for math, ELA, and biology classrooms. The estimates are much less precise given the smaller sample sizes. Only the ELA and biology results are statistically different from zero on their own, but in each of the three, we could not reject that the coefficient on the teacher effectiveness estimate was one.

VII. Conclusion

To develop, retain, and reward great teachers, schools systems must be able to know how to recognize effective teaching. In pursuit of that goal, schools have begun to provide more differentiated feedback to teachers using student achievement gains, classroom observations, and student surveys. Yet there have been legitimate concerns raised about the validity of the measures being used (Rothstein, 2010; Baker et al., 2010; Darling-Hammond et al., 2012).

In this report, we have attempted to provide answers—at least to some of those questions. After randomly assigning seemingly more and less effective teachers to different classrooms, we found that a composite measure of effectiveness (with appropriate controls for prior student achievement) can identify teachers who produce higher achievement among their students. Moreover, the actual impacts on student achievement (within a school, grade, and subject) were approximately equal on average to what the existing measures of effectiveness had predicted. These are causal impacts, estimated with random assignment. In other words, not only do the measures seem to identify more effective teachers, the average impact of being assigned a more effective teacher aligns with expectations. In addition, the teachers who were identified as being effective in raising achievement on state tests had positive effects on other tests as well—although the magnitude of the impacts were only two-thirds as large.

Nevertheless, there were many questions that we could not answer. For instance, we cannot say whether the measures perform as well when comparing the average effectiveness of teachers in different schools. It is a legitimate question, since the measures are being used not only to compare teachers within schools but also to draw inferences about effectiveness between schools. However, we were not able to provide an answer, given the obvious difficulties in randomly assigning teachers or students to different schools.

The current measures—while correctly predicting the causal teacher impacts on student test scores on average—are prone to substantial error. Findings in a companion paper from the Measures of Effective Teaching project (Mihaly et al., 2013) suggest that, for a typical teacher, one year of data on value-added for state tests is highly correlated with a teacher's stable impact on student achievement gains on state tests. However, the measures of value-added on state tests are better at identifying the teachers who will promote gains on current state tests than on alternative assessments: Correlations with a teacher's underlying impact on the other assessments were about half as large. In addition, there is still considerable room for improvement in the measures, especially in English. That improvement could come from better assessments (especially in literacy), more reliable and discerning classroom observations, better student surveys, or measures we did not test here.

Of course, the degree of bias or the amount of error could also worsen as the measures are implemented for high-stakes purposes. It is unrealistic to do a random assignment validity study such as this every year or to validate every new measure of teaching. There is a good reason why this was the first large scale effort to test the validity of the measures with random assignment, despite decades of non-experimental evidence suggesting the importance of teachers. Nevertheless, these results should be updated at some point in the coming years as stakes are attached.

Overall, our findings suggest that existing measures of teacher effectiveness provide important and useful information on the causal effects that teachers have on their students' outcomes. No information is perfect, but better information should lead to better personnel decisions and better feedback to teachers.

References

Baker, E.L., et al. (2010). "Problems with the use of student test scores to evaluate teachers." Economic Policy Institute briefing paper no. 278.

Castellano, K.E., & Ho, A.D. (under review). "Aggregate-level conditional status metrics: From median student 'growth' percentiles to 'value-added' models." Harvard Graduate School of Education working paper.

Chamberlain, G., & Imbens, G. (2004). "Random effects estimators with many instrumental variables." *Econometrica 72*(1): 295–306.

Chetty, R., Friedman, J., & Rockoff, J. (2011). "The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood." National Bureau of Economic Research working paper no. 17699.

Conover, W. (1999). *Practical Nonparametric Statistics (Volume 3).* New York: Wiley.

Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., et al. (2012). "Evaluating teacher evaluation," *Phi Delta Kappan 93*(6): 8–15

Ehlert, M., et al. (under review). "Selecting growth measures for school and teacher evaluations." University of Missouri working paper.

Ferguson, R. (2009). "Tripod student survey, MET project upper elementary and MET project secondary versions." Distributed by Cambridge Education, Westwood, MA.

Goldhaber, D., Walch, J., & Gabele, B. (2012). "Does the model matter? Exploring the relationship between different student achievement-based teacher assessments." University of Washington-Bothell, Center for Education Data and Research working paper no. 2012-6, August 10.

Gordon, R., Kane, T.J., & Staiger, D.O. (2006). "Identifying effective teachers using performance on the job." Hamilton Project discussion paper, published by the Brookings Institution, March.

Jacob, B.A., & Levitt, S.D. (2003). "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *Quarterly Journal of Economics 118*(3): 843–877.

Mihaly, K., et al. (2013). "A composite estimator of effective teaching." Technical report for the Measures of Effective Teaching project, January 8.

Kane, T.J., & Staiger, D.O. (2012). "Gathering feedback for teaching." Research report for the Measures of Effective Teaching project, January.

Kane, T.J., & Staiger, D.O. (2010). "Learning about teaching." Research report for the Measures of Effective Teaching project, December.

Mundlak, Y. (1978). "On the pooling of time series and cross section data." *Econometrica 46*(1): 69–85.

Rothstein, J. (2010). "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics 125*(1): 175–214.

Snedecor, G.W., & Cochran, W.G. (1989). *Statistical Methods, 8th Edition.* Ames, IA: Iowa State University Press.

Staiger, D.O., & Stock, J.H. (1997). "Instrumental variables regression with weak instruments." *Econometrica 65*(3): 557–586.

Stock, J.H., Wright, J.H., & Yogo, M. (2002). "A survey of weak instruments and weak identification in generalized method of moments." *Journal of Business & Economic Statistics 20*(4): 518–529.

Tough, P. (2012). *How Children Succeed: Grit, Curiosity, and the Hidden Power of Character.* New York: Houghton Mifflin.

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." *Journal of the American Statistical Association 57:* 348–368.

Appendix A: Description of Randomization Process

The MET project design called for all teachers participating in the study ("MET teachers") to be randomly assigned one class of students for the 2010–11 school year. When schools joined the study during the 2009–10 school year, principals identified groups of teachers in which all teachers met the following criteria:

1. They were teaching the same subject to students in the same grade (for example, teachers teaching math to 6th graders or English language arts to 8th graders or self-contained 4th grade classes);

2. They had the necessary certification so they could all teach common classes; and

3. They were expected to teach the same subject to students in the same grade in the 2010–11 school year.

These groups of teachers were referred to as "exchange groups," and schools needed at least one exchange group with two or more teachers who agreed to enroll in the study to participate in the MET project.[1]

The plan called for identifying one class roster of students for each teacher in an exchange group and randomly assigning these rosters to the exchange group teachers. The randomized rosters would be chosen from classes of the grade-level and subject of the exchange group. For instance, if the common grade-level and subject were 8th grade math when the teacher enrolled, then only rosters for 8th grade math would be part of the randomization. We call the set of rosters that could be randomly assigned to teachers in the exchange group the "exchangeable rosters."

This appendix explains the procedures used to identify the exchangeable rosters and randomly assign rosters to MET project teachers. It also provides summaries of the randomization process.

RANDOMIZATION PROCESS

The randomization process started in early spring 2010 with the MET project gathering information from all of the partner districts on their scheduling procedures and their methods for exchanging information about assignments between schools and the district central office data system. On the basis of these meetings, the project developed a plan in which schools would complete a spreadsheet with the schedule of courses to be taught by exchange group teachers. Schools would complete the spreadsheet as soon as the schedules became available throughout spring and summer 2010. Schedules would typically be ready before the corresponding class rosters were available. Schools would send the schedules to the MET project team by deadlines dictated by timelines established by each district. The MET project team would process the schedules and make random assignments. Again, according to district timelines, districts would send the MET project team the rosters for all the classes on the schedules. When the rosters were received and verified, the MET project team would send the district and schools the teacher assignments according to separate procedures established with

1 The eligibility requirement served as a guideline and was widely but not completely enforced.

each district.[2] The timelines for completing randomization were set by each district's timeline for completing its class assignments and often required MET project to randomize rosters to teachers within a day or two after the deadline for receiving the spreadsheet schedules.

Figure A presents an example of the spreadsheet used by schools to share scheduling information with the MET project staff. The MET project prepared a custom spreadsheet for each school with the first six rows of data filled in, including the school, district, and teacher project or MET project identification numbers, the teachers' names and district identification numbers. Information filled in by the MET project team also included the exchange group identification number, the grade-level of eligible classes for the exchange group, and a subject code for the eligible subject (e.g., four for middle school math). The spreadsheet contained one page for each exchange group and a table of contents listing all the exchange groups.

Figure A

EXAMPLE OF COMPLETED SPREADSHEET TEMPLATE WITH BLOCK AND PERIOD INFORMATION FOR MET PROJECT TEACHERS

					MET Project Teacher ID	MET Project Teacher ID	MET Project Teacher ID	MET Project Teacher ID	MET Project Teacher ID
					10XX10	10XX11	10XX12	10XX13	10XX14
Dist MET Project ID	School MET Project ID	MET Project Exch. Group	MET Project Eligible Grade	MET Project Subject ID	DISTRICT Teacher ID	DISTRICT Teacher ID	DISTRICT Teacher ID	DISTRICT Teacher ID	DISTRICT Teacher ID
1	9999	DG0999	7	4	999905	999904	999903	999902	999901
					MET Project Teacher Name	MET Project Teacher Name	MET Project Teacher Name	MET Project Teacher Name	MET Project Teacher Name
Class Period	Grade Level	Course Name / Type	Course Section No.		Jane Jones	Kate Knudson	Luke Lesser	Mary May	Nate Newcomb
2	7	Standard Plus Math 7	20212000-06/363		X	X			NA
2	7	Standard Plus Math 7	20212000-12/312		X	X			NA
4	7	Standard Plus Math 7	20212000-18/375				X	X	NA
4	7	Standard Plus Math 7	20212000-24/399				X	X	NA

2 This process was not followed in one district. That district centrally managed all scheduling and could produce a data file of assignments. For that district, schools created schedules with preliminary assignments of MET project teachers and entered them into the districtwide scheduling system. Schools also entered the student rosters into a districtwide database. From its scheduling and rostering databases, the district provided the MET project with the scheduling database and rosters. MET project staff identified exchangeable classes for MET project teachers in the database and made the random assignments.

Schools added data on the teachers' schedules for eligible classes (rows 7 to 10 in Figure A). This information included the period of the day in which the class was to occur, the grade-level of the class, course name or type, and a course section number. The school also put an "X" in the rows of the column corresponding to each teacher's name if the teacher's schedule permitted him or her to teach the class during the period listed. The cells in the columns for the teacher were left blank if the teacher's schedule did not allow the teacher to teach the class. The school put "NA" in every row in a column corresponding to a teacher's name if the teacher had left the school, would not be teaching in the school or grade-level and subject in the 2010–11 school year, or requested not to be part of the MET project in year 2. The MET project team included in the spreadsheets every teacher who was participating in the study at the time that the spreadsheet was created.

Schools received detailed written instructions on how to complete the spreadsheets. Project staff also conducted webinar training for school staff on the randomization process, including how to complete the spreadsheet and how and when random assignments would be communicated with the schools. Some schools completed the spreadsheets accurately, but many made errors that project staff had to assist schools in correcting. Some schools never completed the spreadsheet and project staff, including the district liaison (or district project coordinator), needed to call these schools, obtain the information via phone, and complete the spreadsheet.

In the example in Figure A, Jane Jones and Kate Knudsen could both teach either section in period 2, but they could not teach grade 7 math in period 4. Luke Lesser and Mary May were the opposite: They could teach grade 7 math in period 4 but not during period 2. Nate Newcomb would not be teaching grade 7 math at the school in the 2010–11 school year or had decided not to participate in the study in year 2. This situation in which not all the teachers in the exchange group were scheduled to teach during a common period occurred very frequently among participating schools. To accommodate this lack of a common period, the MET project created subgroups within the exchange group of teachers who were scheduled to teach in a common period and could exchange rosters. In the example in Figure A, there would be two subgroups of the exchange group: a period 2 group with Jane Jones and Kate Knudson and a period 4 group with Luke Lesser and Mary May. These subgroups were called "randomization blocks," and rosters were randomly assigned among teachers in the same randomization block. Each teacher could belong to only one randomization block.[3] If teachers were in two or more blocks, they were randomly assigned to a block. For instance, suppose Kate Knudson could also teach in period 4 and Luke Lesser could also teach in period 2. They both would be in two possible

3 In some very rare occasions the following situation occurred:

Class Period	Grade Level	Course Name / Type	Course Section No.	Jane Jones	Kate Knudson	Luke Lesser
2	7	Standard Plus Math 7	20212000-06/363	X	X	X
2	7	Standard Plus Math 7	20212000-12/312	X	X	X
4	7	Standard Plus Math 7	20212000-18/375		X	X

There is one section in period 4 that could be taught by either Knudson or Lesser but not Jones. All three teachers can teach in period 2. There are three teachers and three sections but one teacher is not available to teach one of the sections. In this case, the project first randomly chose between Knudson and Lesser to receive the period 4 roster (say we chose Lesser) and then randomly assigned the period 2 rosters to the other two teachers (Jones and Knudson). We treat Knudson as being in two blocks: one with Jones and one with Lesser, even though Knudson only taught one randomly assigned roster.

randomization blocks and project staff would randomly assigned Knudson to one block and Lesser to the other. If only one teacher was available to teach during a period, the project called that teacher a singleton and that teacher was not randomly assigned a roster.

Within a randomization block, teachers were randomly sorted and rosters (sections) were randomly sorted and the first teacher was matched with the first roster and so on.

RANDOMIZATION SUMMARY

The project requested scheduling information for 2,462 teachers from 865 exchange groups in 316 schools. The project created 668 randomization blocks from 619 exchange groups in 284 of the participating schools. The remaining schools' schedules did not permit randomly swapping rosters among any of MET project teachers or all its MET project teachers had left the school or the study.

From these randomization blocks, the project randomly assigned rosters to 1,591 teachers.[4, 5] (This includes 386 high school teachers and 24 teachers for whom rosters were later found to be invalid.) Seven hundred, seventy teachers were not eligible for randomization because they were not scheduled to teach the exchange group subject and grade level in 2010–11 or they decided not to participate in year 2 of the study. The remaining 281 teachers could not be randomized because they did not teach in a period with two or more teachers for exchanging rosters.

4 Two teachers in blocks with a single teacher were randomly assigned rosters and counted in the randomized sample. These teachers were included in the analysis sample but do not contribute to estimates.

5 Because of a large number of teachers without exchangeable rosters in one district, the study added 33 teachers who did not participate in year 1 to the study and included them in the random assignment of the rosters. The remaining 1,558 teachers with randomly assigned rosters all participated in year 1.

Appendix B: Treatment of Outliers

As described in the text, we dropped teachers whose value-added scores put them in the top 1 percent of all teachers. In Appendix **Table B1**, we report the robustness of the results to a variety of different cut-offs. In addition to reporting the effects with the full sample (first column), the remaining columns of the table report the results of dropping the top .5 percent, top 1 percent (our preferred option), top 2 percent, and top 3 percent. (We were particularly interested in the outliers on the high side, since that's where other researchers had pointed to evidence of malfeasance.) However, abuses in one year could lead to unusually low gains for the teachers who taught the same students in the subsequent year. As a result, we also explored the implications of similar cutoffs for teachers at the bottom of the distribution.

Table B1

ROBUSTNESS TO DIFFERENT DATA TRIMMING

		Drop block if teacher value-added estimate is in the top ...				Drop block if teacher value-added estimate is in the top or bottom ...			
Variables	**Full Sample**	**0.50%**	**1%**	**2%**	**3%**	**0.50%**	**1%**	**2%**	**3%**
Math and ELA Stacked:									
Expected student achievement based on teacher's effectiveness	0.811*** (0.146)	0.944*** (0.119)	0.955*** (0.123)	0.947*** (0.127)	0.950*** (0.126)	1.004*** (0.120)	1.041*** (0.124)	1.017*** (0.136)	0.919*** (0.136)
Observations	27,790	27,677	27,255	26,571	25,845	27,448	26,816	25,795	24,327
R-squared	0.682	0.683	0.684	0.683	0.681	0.683	0.685	0.684	0.678
Math:									
Expected student achievement based on teacher's effectiveness	0.996*** (0.132)	1.031*** (0.133)	1.039*** (0.138)	1.048*** (0.144)	1.070*** (0.143)	1.107*** (0.132)	1.135*** (0.137)	1.136*** (0.149)	1.057*** (0.148)
Observations	13,714	13,672	13,457	13,111	12,765	13,480	13,122	12,637	11,922
R-squared	0.704	0.704	0.705	0.703	0.701	0.705	0.706	0.705	0.699
English Language Arts:									
Expected student achievement based on teacher's effectiveness	0.375 (0.264)	0.677*** (0.203)	0.697*** (0.213)	0.631*** (0.216)	0.637*** (0.214)	0.682*** (0.203)	0.727*** (0.212)	0.665*** (0.233)	0.654*** (0.237)
Observations	14,076	14,005	13,798	13,460	13,080	13,968	13,694	13,158	12,405
R-squared	0.664	0.664	0.666	0.667	0.663	0.665	0.667	0.666	0.661

Note: The sample consists of all randomized students in grades 4 through 8. The dependent variable is student achievement on state tests following random assignment in 2011, standardized by grade and district. Expected student achievement in a teacher's class is the prediction of the teacher's value-added in that subject, based on value-added, student surveys, and observations in the prior school year. The coefficients are LIML IV estimates of the effect of actual teacher effectiveness in the relevant subject, using the effectiveness of randomly assigned teachers interacted with school-level indicators as instruments. All specifications also included fixed effects for randomization block, and controls for students' prior achievement and demographics as described in the text. Standard errors were calculated allowing for clustering within teacher.

The top panel reports the results when math and ELA results are combined, and students' scores in math and ELA are stacked and treated as separate observations (although not assumed to be independent, since we are also allowing for clustering of errors at the school level). In the full sample, the coefficient on expected teacher effectiveness is .811 with a standard error of .146. In the remaining columns reporting the results of different rules for trimming outliers, the coefficients ranged from a low of .919 to a high of 1.041. The standard errors ranged from .12 to .14.

The next panel reports the results for math. The coefficient estimated using the full sample is .996 with a standard error of .132. The coefficients estimated with different amounts of trimming ranged from 1.031 to 1.136, with standard errors ranging from .13 to .15.

The bottom panel reports the results for ELA. Although the trimmed results are all statistically different from zero and not statistically different from one, the coefficients range from a low of .631 to a high of .727, with standard errors ranging from .20 to .24. The only coefficient that is not statistically different from zero is the coefficient on the full sample in ELA, where the coefficient is .375, with a standard error is .264.

In the ELA classrooms, there was one exchange group with one teacher who had an extraordinarily large estimated value-added. Appendix **Figure B1** reports the difference in teachers' estimated value-added (or adjusted student achievement gain) in 2009–10 relative to the average in their exchange group. The teachers taught 5th grade in elementary school and had a difference of more than 1.5 student-level standard deviations in their value-added estimates in ELA. (One teacher was .8 standard deviations better than the average and the other teacher was .8 standard deviations worse that the average.) The difference in value-added within this one pair of teachers was equivalent to 1.6 teacher-level standard deviations.

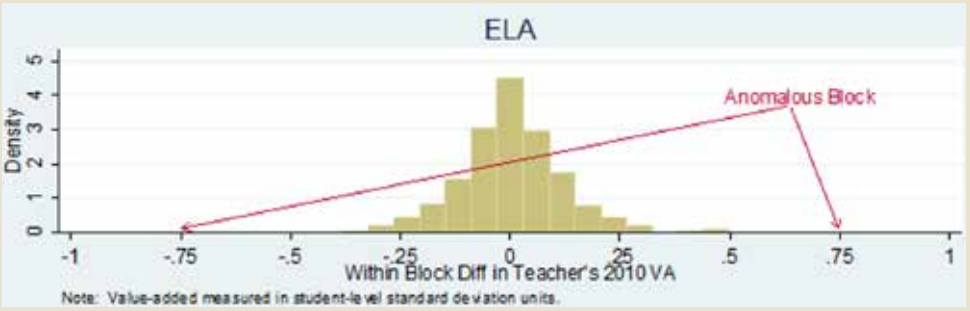
Moreover, there is reason to believe that the outcome data for this pair of teachers are incorrect. Appendix **Table B2** reports various measures for these two teachers, referred to as Teacher A and Teacher B. Teacher B started out with higher estimated value-added in 2008–09—.494 versus .040—resulting in a difference of .45. However, the difference more than tripled between 2008–09 and 2009–10—1.493 versus -.173—resulting in a difference of 1.666 student-level standard deviations. (This is almost twice as large as the black-white achievement gap among 8th graders nationally.)

The next panel of Appendix Table B2 reports the scores over time for the cohort of 5th graders in 2009–10 in these two teachers' classrooms. The students assigned to Teacher A were always above average for their grade level and district. In 2008, they were .038 standard deviations above the grade mean in ELA in their district. In 2010, their scores were .304 standard deviations above the grade-level mean. In 2011, they maintained their position at .296 standard deviations above in 2011. However, the students assigned to Teacher B showed a very different pattern: In 2010, they scored 1.459 standard deviations *above* the mean and then observed a large decline in performance, scoring .483 standard deviations *below* the mean in 2011. A swing of 2 standard deviations for a classroom of youth is unusual in a single year.

The last panel reports the mean scores of the 2009–10 cohort on the supplemental assessments administered in the MET project. In contrast to their state ELA scores, the students in Teacher B's class had scores close to the mean (.16 standard deviations higher than the mean on the SAT9 OE and .02 standard deviations below the mean on the Balanced Assessment in Mathematics) and lower than the scores for the students in Teacher A's class.

Figure B1

DIFFERENCE IN TEACHER VALUE-ADDED RELATIVE TO THE BLOCK MEAN



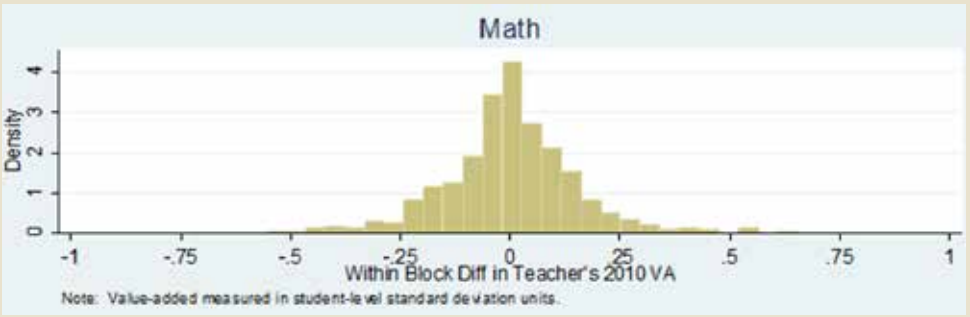


Table B2

SUMMARIZING DATA FOR ANOMALOUS RANDOMIZATION BLOCK

	Teacher A	Teacher B
Estimated Value-Added (ELA)		
2008–09	0.040	0.494
2009–10	-0.173	1.493
State ELA Scores for 2009–10 Cohort of 5th Grade Students		
2008	0.038	-0.335
2009	0.623	0.237
2010	0.304	1.459
2011	0.296	-0.483
MET Project Supplemental Assessments in 2010		
BAM	0.389	-0.020
SAT9	0.374	0.163

Note: The above block was dropped from the analysis.

Bill & Melinda Gates Foundation

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

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

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org