Educational Evaluation and Policy Analysis
Month 201X, Vol. XX, No. X, pp. 1–19
DOI: 10.3102/0162373714537551

Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence From the Randomization of Teachers to Students

Rachel Garrett
American Institutes for Research

Matthew P. Steinberg
University of Pennsylvania

Despite policy efforts to encourage multiple measures of performance in newly developing teacher evaluation systems, practical constraints often result in evaluations based predominantly on formal classroom observations. Yet there is limited knowledge of how these observational measures relate to student achievement. This article leverages the random assignment of teachers to classrooms from the Measures of Effective Teaching (MET) study to identify teacher effectiveness using scores from the Framework for Teaching (FFT) instrument, one of the most widely used classroom observation protocols. While our evidence suggests that teacher performance, as measured by the FFT, is correlated with student achievement, noncompliance with randomization and the modest year-to-year correlation of a teacher's FFT scores constrain our ability to causally identify effective teachers. Implications for policy and practice are discussed.

Keywords: *education policy, teacher effectiveness, teacher evaluation, Framework for Teaching (FFT)*

Introduction

Education policymakers have long recognized the importance of having highly effective teachers in all of our nation's classrooms. Among recent policy efforts, the federal Race to the Top (RTTT) competition and the No Child Left Behind (NCLB) waivers have highlighted the importance of ensuring highly effective instruction as a central policy consideration, primarily through increased attention to teacher evaluation systems. This policy focus follows from research findings that have emphasized the importance of teacher quality for student achievement above other school-level characteristics (Aaronson, Barrow, & Sander, 2007; Goldhaber, 2002; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004).

Even as federal policy efforts have promoted more rigorous approaches to evaluating teacher performance, little consensus exists about the most salient measures of teacher effectiveness for the purposes of teacher accountability (such as high-stakes tenure decisions) and compensation (such as merit-based pay programs). Traditional measures of teacher performance, such as years of experience or educational attainment, play little role in student achievement (see Steinberg & Sartain, in press, for a summary). Newer efforts incorporate student test score data through value-added measures of teacher performance, yet such measures can be statistically unreliable over time (Goldhaber & Hansen, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009), and prone to severe bias due to the nonrandom assignment of students

to teachers (Rothstein, 2009). Beyond these concerns related to value-added scores, the nonrandom assignment of students to teachers has long been acknowledged (Monk, 1987), and typically results in the systematic placement of higher performing students in the classrooms of higher performing teachers (Kalogrides, Loeb, & Beteille, 2013). Given this nonrandom sorting of students to teachers, it becomes increasingly difficult to isolate a teacher's unique contribution to student learning.

In the wake of RTTT's teacher evaluation mandate, states and local districts have incorporated three primary measures of teacher performance into their teacher evaluation systems. Specifically, teachers are being assessed using ratings from protocols for classroom-based observations, value-added scores, and student learning objectives (SLOs), which may be either state or locally determined. When multiple measures are incorporated into teacher performance evaluations, teacher effectiveness has been shown to be more reliably measured (Kane & Staiger, 2012), while also signaling to educators that the multifaceted complexity of their jobs is taken into account when assessing their performance. In practice, however, the incorporation of multiple measures is limited. For example, more than two thirds of all teachers nationwide teach in grades or subjects that are not tested by state-mandated achievement exams, and therefore value-added estimates of their performance are unavailable (Watson, Kraemer, & Thorn, 2009). Even when test scores are available, teachers, principals, and unions have been particularly concerned with the use of value-added measures in teacher evaluation, stemming from factors including the year-to-year volatility, the opaque nature of how these measures are constructed, and their inability to provide instructional guidance to help teachers improve both their practice and student achievement. The implementation of SLOs for understanding teacher performance is also tenuous, as many jurisdictions are currently in the process of creating the SLOs, and determinations for how to use SLOs in evaluation systems are in the very early stages.

In contrast, classroom observations occupy a different place in the evolving teacher evaluation landscape. Historically, the entirety of a teacher's summative, year-end evaluation was based on principal observation of classroom practice. As classroom observations have long been the hallmark of teacher evaluations, teachers are familiar with the process and may feel more comfortable with it. As opposed to value-added scores generated by a complex statistical process, scores from classroom observation protocols are more straightforward and transparent for educators to connect to their actual work, thereby enabling various stakeholders to feel more confident about the use of classroom observation protocols for teacher performance evaluation.

As a result, observational measures of teacher instructional practice have emerged as critically important components of teacher ratings. Among the observational protocols currently available, one of the most commonly used (Brandt, Mathers, Oliva, Brown-Sims, & Hess, 2007) is the Danielson Framework for Teaching (FFT; Danielson, 1996). The FFT rates teachers at four levels of performance—unsatisfactory, basic, proficient, or distinguished—and includes four domains of teacher effectiveness that capture planning and preparation, classroom environment, instruction, and professional responsibilities. It is meant to capture key features of effective teaching, irrespective of grade or subject, making it applicable to all teachers. As a general observational protocol, the FFT may be the only measure used in the evaluation of teachers in untested grades and subjects, and may serve as the only option for a consistent measure of teacher performance across all teachers. In cases where student test scores are unavailable and SLOs are not yet in place, the FFT may account for upward of 85% of a teacher's performance evaluation.[1]

However, there are unanswered questions about the use of the FFT in evaluation systems. As with other measures of teacher performance, the nonrandom assignment of students to teachers has limited researchers' ability to understand the extent to which a teacher's performance, as measured by the FFT, relates to student academic achievement. Furthermore, while composite measures of teacher performance that include FFT scores (along with student surveys and value-added estimates) have been found to be positively associated with student test performance (Kane, McCaffrey, Miller, & Staiger, 2013), a more thorough understanding of whether FFT measures

alone can identify teacher effectiveness is fundamental to informing policy-driven teacher evaluation systems. Considering that the FFT is being adopted or adapted widely across the country as the rubric for evaluating classroom teaching, it is critical to understand whether we are able to identify effective teachers based solely on FFT classroom observation scores.

This study aims to provide information to improve this understanding. Using data from the Measures of Effective Teaching (MET) study, we analyze the key components of the FFT and then link those components to student performance. We augment the evidence base by using the randomization of teachers to classrooms to estimate teacher effectiveness using multiple years of observation scores from the FFT. We find that a simple composite of the eight components of the FFT best characterizes a teacher's observed instructional practice. Using this measure, we find that both reading and math instruction are highly correlated with student achievement; however, these associations are prone to bias induced by student sorting postrandomization. After accounting for the nonrandom sorting of students to teachers, we are unable to identify teacher effectiveness using just 2 years of classroom observation scores, due in large part to the relatively modest year-to-year correlation of a teacher's FFT scores.

We proceed by discussing the role of classroom observation protocols in educator effectiveness systems and the existing evidence on the FFT. Next, we describe the data and the empirical methods used to estimate teacher effectiveness using classroom observation scores from the FFT. We then present and discuss the findings, attending to the policy implications of our results for evolving teacher evaluation systems.

Background and Previous Research on the FFT

In the teacher evaluation process, the observation-based component of teacher performance aims to provide both instructional support to teachers and a means of assessing their performance. In principle, the efficacy of observation protocols, such as the FFT, rests in the conferences between teachers and classroom observers, typically the school's principal. This feedback mechanism is a key feature of the FFT, whereby practitioners and their observers use the protocol to structure conversations about ways to improve instructional practice. Ideally, a practitioner would work with an observer several times over the academic year and engage in an ongoing dialogue shortly after each observation, which can lead to a heightened awareness of current practices and even targeted professional development opportunities. Ultimately, the FFT is used to give teachers both the opportunity and support to improve their practices, as well as for evaluation and accountability purposes. As such, the use of an observational protocol like the FFT can embody both "the carrot and the stick" for teachers within an evaluation system.

In the absence of either this structured feedback or the accountability for ratings, the relationship between observed instructional practices and student learning captures the "status quo," in which teachers optimize their practices based on private (and likely imperfect) information about their performance. In the MET study, teacher observations were conducted remotely, absent the conference component of the FFT or the provision of official performance ratings. As a result, this study does not allow for an understanding of how teacher instructional practice may affect student performance when being directly targeted for manipulation, via the feedback loop, as intended by FFT.[2]

Despite the widespread use of the FFT, there is limited research studying its use, and most is observational. Descriptive studies have found small to moderate positive correlations of FFT scores with student learning, with some variation by grade and subject (Gallagher, 2004; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004). While these studies provide basic descriptions of the associations between FFT measures and student achievement, they do not shed light on the extent to which teacher effectiveness may be identified using classroom observation scores from the FFT, and thus they do not provide the evidence needed to justify the FFT's use in policy and teacher evaluation systems. More recent research from Cincinnati Public Schools suggests that teacher observation and evaluation using the FFT promotes student achievement growth in math both during the school year in which the teacher is evaluated as

well as in the years after evaluation; it is important to note, however, that the researchers found less consistent impacts on reading achievement (Kane, Taylor, Tyler, & Wooten, 2011; Taylor & Tyler, 2012). Kane et al. (2011) provide a more nuanced understanding of the instructional practices that may drive the observed improvements to student achievement through a principal components analysis (PCA). The authors use three different component groupings to understand the relationship with student achievement, but find the most consistent, significant results from the component that includes a fairly equal, positive weighting of all the available FFT components. Further work is needed to test if the results of the PCA would replicate consistently over a different sample with potentially different underlying variation. Also, the authors indicate that the observed relationships between FFT scores and student achievement may still be biased due to the nonrandom assignment of teachers to students.

Two studies that consider the FFT provide causal evidence. First, in research studying a randomized pilot using the FFT for teacher evaluation among elementary schools in the Chicago Public School district, Steinberg and Sartain (in press) find that schools randomly assigned to participate in the teacher evaluation pilot realized significant improvements in reading performance, and positive but not significant improvements in math, relative to student achievement in schools that were not randomly assigned to use the FFT. Using the randomized sample in the MET data, Kane et al. (2013) are able to identify teacher effectiveness using a composite measure of teacher performance that includes FFT scores, student survey responses, and teacher value-added scores. Yet the goal of Kane et al. was to document how well multiple measures of teacher performance together capture teacher effectiveness, rather than to specifically gauge whether teacher effectiveness could be identified using FFT scores alone.

Given the policy focus that is mobilizing widespread use of the FFT for teacher evaluation purposes, and the reality that the ability to assess multiple measures of teacher performance is often constrained, researchers and educators alike need clear evidence to understand how FFT scores function independently in capturing the role of teachers in student achievement.

Data and Sample

To address our research aim, we use newly available data from the MET study. The MET study was carried out over 2 school years (2009–2010 and 2010–2011 years) and across six districts, with the goal of enabling the study of how to measure and empirically discern high-quality instruction. The six participating districts were Charlotte-Mecklenburg (North Carolina) Schools, Dallas (Texas) Independent School District, Denver (Colorado) Public Schools, Hillsborough County (Florida) Public Schools, Memphis (Tennessee) City Schools, and the New York City (New York) Department of Education (White & Rowan, 2012). In the summer of 2010, just prior to the MET study's second academic year, a subsample of classes of students were randomly assigned to teachers. The randomization sample consisted of 1,559 teachers in 284 schools in six districts (White & Rowan, 2012). The randomization was conducted within randomization blocks that consisted of grade–subject combinations of classes within a school. Importantly, at least two teachers must have been teaching in the same grade–subject combination (e.g., seventh-grade mathematics) to be included in the randomization sample.

We observe information on student test scores on state-mandated exams for the first 2 years of the study (i.e., 2009–2010 and 2010–2011) and up to 3 years prior. For our sample of students in Grades 4 to 8, we focus on math and reading test scores as our outcome measures, which are standardized (z scores) within district, subject, and grade. We supplement the student test score data with student-level demographic characteristics, including a student's race/ethnicity, gender, age, special education status, free-lunch status (as a proxy for student poverty), gifted status, and whether a student is an English language learner (ELL). We also observe teacher background characteristics, including gender, race/ethnicity, degree status, and years of teaching experience in the district.

To assess the extent of student noncompliance with their random teacher assignment, we create an indicator for whether the student's actual teacher was the teacher to whom the student was randomly assigned. Given the extent of noncompliance across school districts in the study (to be

discussed later in the article), we use this variable to explore characteristics related to noncompliance. Specifically, as the randomization of teachers to classes should lead to an equal distribution of student and teacher characteristics, we use the student and teacher information to understand the patterns of selection that occur when the randomization did not hold due to noncompliance.

Measure of Teacher Instructional Practice: FFT

We use the FFT to measure a teacher's instructional practice. The MET data include eight components from two of the FFT domains, capturing Classroom Environment and Instruction.[3] For each section (class) that a teacher was responsible for (e.g., fifth-grade math), one external rater generated FFT scores for each of the eight components from a single segment (i.e., a 30–35 minute lesson conducted by the teacher that the external rater observed via video recording of the lesson).[4] The FFT rating scores for each component (by subject—English language arts [ELA] or math) were then averaged (using the harmonic mean) across all segments to create section-level aggregate scores. These aggregate scores represent the average FFT scores for a teacher's subject-specific instruction for a single section of students. The average section-specific FFT scores are the primary measures of a teacher's observed effectiveness we use to construct the covariate of interest in our study (we refer to this covariate as *Teach* and discuss its construction later in the article).

Study Sample

We focus on students and teachers in Grades 4 to 8 who were part of the randomization sample. In addition, as we focus our attention on teacher effectiveness in a given subject area, we created two analytic samples: one to estimate teacher effectiveness in reading (ELA) and a second to estimate teacher effectiveness in math. For students in a section randomly assigned to a MET teacher, we included those from fourth to eighth grade who had complete test score data for both the 2009–2010 and 2010–2011 school years for the subject being studied in the analytic sample. Sample inclusion also required that a student's actual and randomly assigned teachers have subject-specific FFT score information available for the second study year; for the randomly assigned teacher, we also required subject-specific FFT score information for the first study year. While there was virtually no missing data on student characteristics included as covariates in the analysis,[5] there was some missingness among the teacher characteristics. For the ELA analytic sample, the missingness ranged from 3.8% (gender and race variables) to 22.9% for degree status and 26.9% for total years of teaching in the district. For the math analytic sample, the missingness ranged from 2.8% (gender variables) to 21.8% for years teaching in the district and 24.8% for degree status. To retain these observations, we assigned a value of 0 to missing values and created indicator variables for missing in the analyses.

Table 1 summarizes the student-level analytic samples by subject area. Our final analytical sample for ELA includes 8,780 students, with 581 teachers in 163 schools, whereas our math sample has 7,934 students with 528 teachers in 156 schools. The two samples are very similar in terms of student, teacher, and classroom composition characteristics. The average age of students in the analytic samples is approximately 10 years old, 49% of students are male, 7% of students receive special education services, approximately 15% are ELL, and approximately 60% receive free or reduced-price lunch. Moreover, approximately 25% of students are White, approximately 30% are Black, approximately 35% are Hispanic, and 8% are Asian. The only notable difference between the ELA and math analytic samples is that students in the ELA sample are twice as likely to be academically gifted than those in the math sample (12% vs. 6%) and, correspondingly, are in classes with a higher percentage of gifted children. Also, while the average age is comparable, students in the ELA sample are more likely to be distributed in the higher grades (58% of the ELA sample are in fourth or fifth grade compared with 64% of the math sample).

Empirical Approach

Capturing Teacher Instructional Practice

A key empirical consideration relates to modeling a teacher's instructional practice, which we

TABLE 1
Summary Statistics for Analytic Samples

	ELA sample	Math sample
Student characteristics		
Age	10.4 (1.5)	10.1 (1.4)
Fourth grade	26%	31%
Fifth grade	32%	33%
Sixth grade	15%	16%
Seventh grade	13%	10%
Eighth grade	13%	11%
Male	49%	49%
Gifted	12%	6%
Special education	7%	7%
ELL	13%	15%
Free/reduced-price lunch	59%	61%
White	27%	25%
Black	28%	29%
Hispanic	34%	35%
Asian	8%	8%
Other race	2%	3%
Subject test score 2010	0.17 (0.93)	0.13 (0.92)
Subject test score 2011	0.17 (0.91)	0.11 (0.90)
Teacher characteristics		
Male	13%	15%
White	64%	64%
Black	28%	27%
Hispanic	6%	7%
Other race	1%	2%
Years in district	7.4 (6.4)	7.4 (6.9)
Master's or higher	36%	37%
Classroom characteristics		
Class size	25.9 (5.8)	25.8 (6.6)
Gifted	11%	6%
Male	50%	50%
Special education	8%	8%
ELL	14%	15%
Average age	10.4 (1.4)	10.2 (1.3)
Free/reduced-price lunch	59%	60%
Other race	2%	3%
Asian	8%	8%
Hispanic	33%	34%
Black	29%	30%
White	27%	25%
Average subject test score 2010	0.13 (0.53)	0.10 (0.49)
Students	8,780	7,934
Teachers	581	528
Schools	163	156

Note. Authors' calculations from MET data from the 2010–2011 school year. Mean (and standard deviation) reported. The ELA and math samples consist of all students with complete test score data from both the 2009–2010 and 2010–2011 school years. Teacher characteristics are for a student's actual subject-specific teacher during the 2010–2011 school year. Test scores have been standardized within district, subject, and grade. ELL = English language learner; MET = Measures of Effective Teaching.

refer to as *Teach*. We follow Kane et al. (2011) and conduct a PCA to uncover the extent and nature of the systematic relationships among the eight components of the FFT protocol on which teachers' subject-specific, section-level aggregate scores were rated by external evaluators. As discussed in Kane et al., the eight components of the FFT protocol are highly correlated, and simply including these measures in a regression will generate noisy (and potentially unstable) estimates of the relationship between teacher practice and student achievement. Instead, the composite measures generated by a PCA will be uncorrelated by construction, and will provide clearer insight into the relationship between teacher FFT scores and student achievement. Table 2 summarizes the results of the PCA. While we considered math and ELA separately, the results are highly consistent across both subjects. For both content areas, the first principal component (*Comp1*) explains approximately 65% of the unique variation in the eight FFT components.

We tested the sensitivity of these results by conducting further PCA on different samples within the data. To be more inclusive, we ran a PCA on the full sample rather than the randomization sample, including all grades in the data. We also tested the results on subsamples, based on teacher characteristics (race, etc.), by grade level, and by student characteristics. In all iterations, the results were indistinguishable. Moreover, each criterion for determining the number of principal components to retain—keep any principal component with corresponding eigenvalue greater than 1 (Kaiser criterion) or keep any principal component such that the cumulative variance explained is at least 60% to 70% of the total—led us to the same conclusion: The first principal component best characterized the measure of observed teaching practice (*Teach*). Furthermore, the eight FFT components each explain approximately the same amount of variation in the first principal component, as captured by the eigenvectors. Given both the robustness of the PCA results to alternative sample choices as well as the approximately equal weighting of the eight FFT components, we concluded that the policy-relevant measure of *Teach* was an equally weighted average of all eight FFT components, thereby creating a single index for *Teach*.[6]

TABLE 2
Principal Components Analysis (PCA)

	ELA				Math			
Component	Eigenvalue	Difference	Proportion	Cumulative	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.32	4.44	.6648	.6648	5.17	4.21	.6458	.6458
Comp2	.879	.511	.1099	.7747	.956	.521	.1197	.7655
Comp3	.367	.014	.0459	.8206	.436	.041	.0545	.8201
Comp4	.353	.028	.0441	.8647	.395	.079	.0494	.8694
Comp5	.324	.029	.0406	.9053	.316	.017	.0395	.9090
Comp6	.296	.051	.0369	.9423	.299	.065	.0375	.9464
Comp7	.244	.027	.0305	.9728	.234	.040	.0293	.9757
Comp8	.218	—	.0272	1.000	.194	—	.0243	1.000

	Principal components (eigenvectors)							
FFT component	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4
2a	.3518	.3640	.1236	−.5663	.3630	.3091	−.1233	−.3454
2b	.3751	−.1308	−.2462	−.2852	.3797	−.0966	−.5134	−.1905
2c	.3333	.4552	.0831	.7307	.3395	.4315	.2920	.3976
2d	.3325	.5395	−.2619	−.1080	.3276	.5555	−.1252	.2250
3a	.3590	−.1551	.7843	−.0595	.3563	−.0758	.5674	−.6358
3b	.3539	−.3457	.1611	.1214	.3432	−.3778	.3913	.3297
3c	.3670	−.3019	−.3652	.0350	.3722	−.2798	−.3790	−.0458
3d	.3538	−.3427	−.2723	.1814	.3438	−.4166	−.0402	.3473

Note. The ELA and math results are based on 581 and 528 teachers, respectively, in Grades 4 to 8 in the 2010–2011 school year. These teachers represent the analytic sample of students' actual teachers during the 2010–2011 school year. FFT component 2a corresponds to the "Creating an Environment of Respect and Rapport" component of domain 2, *The Classroom Environment*; 2b corresponds to the "Establishing a Culture for Learning" component of domain 2; 2c corresponds to the "Managing Classroom Procedures" component of domain 2; and 2d corresponds to the "Managing Student Behavior" component of domain 2. FFT Component 3a corresponds to the "Communicating With Students" component of domain 3, *Instruction*; 3b corresponds to the "Using Questioning and Discussion Techniques" component of domain 3; 3c corresponds to the "Engaging Students in Learning" component of domain 3; and 3d corresponds to the "Using Assessment in Instruction" component of domain 3. FFT = Framework for Teaching.

Checking for Balance on Student Characteristics After Randomization

We next check for covariate balance to assess whether the randomization of teachers produced equivalence across classes. To do so, we regress student characteristics on the randomly assigned teacher's measure of *Teach*, controlling for randomization block fixed effects. We then assess whether the coefficient is statistically different from 0. This enables us to test for whether student characteristics are systematically related to a student's randomly assigned teacher's observed measures of instructional practice. In principle, if the randomization produced statistical equivalence, we should not see any statistically significant association between student characteristics and the instructional practice of the randomly assigned teacher. Formally, we estimate the following model to assess covariate balance:

$$X_i^l = \gamma_0 + \gamma_1 (Teach_{ik}) + \theta_r + \mu_i, \quad (1)$$

where $l = 1 \ldots L$ student characteristics, $Teach_{ik}$ is the prerandomization, subject-specific measure of observed instructional practice during the 2009–2010 school year for student i's randomly assigned teacher k, and θ_r is the randomization block fixed effect. We estimate Equation 1 for both the ELA and math analytic samples. Table 3 summarizes the extent to which the randomization of classes to teachers generated covariate balance on the

TABLE 3
Covariate Balance: Student Characteristics and Randomized Teacher Instructional Practice

Student characteristic	ELA sample	Math sample
Age	−.021 (.030)	−.001 (.031)
Male	−.002 (.026)	.029 (.030)
Gifted	.013 (.034)	−.016 (.020)
Free/reduced-price lunch	−.005 (.021)	−.000 (.033)
Special education	−.012 (.018)	.004 (.019)
ELL	−.004 (.031)	−.035 (.030)
Prior (2010) test score	−.010 (.080)	.082 (.072)
White	.001 (.023)	−.001 (.024)
Black	−.034 (.024)	.009 (.022)
Hispanic	.016 (.025)	−.057* (.030)
Asian	.013 (.020)	.033 (.025)
Students	8,780	7,934
Randomized teachers	593	537
Randomization blocks	259	246

Note. Each cell represents a separate regression of an individual student characteristic on the student's randomly assigned teacher's 2010 subject-specific measure of *Teach* (the mean value of the eight FFT components). All regressions control for randomization block fixed effects. Robust standard errors (clustered at the randomization block level) reported in parentheses. ELL = English language learner; FFT = Framework for Teaching. Coefficients statistically significant at the *10%, **5%, and ***1% levels.

observable dimensions of students. For both the ELA and math analytic samples, there is strong evidence that the randomization produced statistical equivalence on student characteristics. For the ELA sample, no individual student characteristic is statistically associated with *Teach*. For the math analytic sample, only the share of Hispanic students is associated with the randomly assigned teacher's 2010 measure of *Teach*, and the association is only marginally significant (at the 10% level). From this, we can conclude that the randomization process was correctly executed.

Estimating Teacher Effectiveness With Classroom Observation Scores

The normal practice by which students are matched to classroom teachers presents a significant obstacle to identifying effective teachers using measures of classroom instruction. In particular, the concern is that assortative matching—where students are positively sorted into classrooms with higher performing teachers—makes it difficult to ascertain if student achievement is being driven by the instruction of the teacher or the background characteristics of the students. One way to disentangle these effects to generate an unbiased estimate of the teacher on student achievement is to randomly assign teachers to classes of students, so that the background characteristics of students and teachers are unrelated to the student–teacher matching in classrooms.

Having verified (above) that the randomization worked, we aim to improve upon previous observational research on the FFT by leveraging the MET study's randomization of teachers to classrooms to estimate a teacher's causal impact on student academic achievement. When classes of students are randomly assigned to teachers, any components of teacher quality—both those components observed using the FFT protocol and unmeasured components—are, by construction, uncorrelated with the characteristics of their classes that also predict academic performance. This is the key identifying assumption underlying the randomization of classes to teachers, and it is this cross-teacher variation that we employ to estimate teacher effectiveness using FFT scores.

To identify teacher effectiveness, we estimate variants of the following value-added model:

$$Achieve_{ijt} = \beta_0 + \beta_1\left(Teach_{jt}\right) + \beta_2\left(Achieve_{i,t-1}\right) + X_{it}\Gamma + Z_{jt}\varphi + \theta_r + \varepsilon_{ijt}, \quad (2)$$

where $Achieve_{ijt}$ is the achievement (math or reading) for student i with teacher j in year t; $Teach_{jt}$ is a measure of observed teaching practice for teacher j in year t; $Achieve_{i,t-1}$ is the achievement (math or reading) for student i in the prior school year ($t - 1$); X_{it} represents observed student characteristics for student i in year t, including race/ethnicity, gender, age, special education status, free/reduced-price lunch status, gifted status, and ELL status; Z_{jt} are teacher characteristics for teacher j in year t, including race, gender, years of experience in the district, and educational attainment (master's degree and

higher); and ε_{ijt} is a random error term. Importantly, as the randomization occurred within randomization blocks, θ_r is a randomization block fixed effect and accounts for the block structure of teacher randomization.

We first estimate Equation 2 using the student's randomly assigned teacher's measure of *Teach*. Estimates of β_1 will produce the intent-to-treat (ITT) effect of assigned teacher quality, as measured by the FFT, on student achievement. To generate the ITT effect, we use the randomly assigned teacher's measure of *Teach* from the 2009–2010 school year (2010 *Teach*). We use this measure instead of the 2011 measure for two reasons. First, the best estimate of the effectiveness of the randomly assigned teacher *at the time of randomization* is the 2010 measure, which represents the intended teacher quality to which students were randomly assigned. Second, we wish to avoid any influence that nonrandom sorting, postrandomization, may have on the 2011 *Teach* measure. Moreover, as the 2011 *Teach* measure is a function of FFT scores based on multiple lesson-specific observations occurring over the course of the school year, this time-varying treatment, if influenced by sorting, is unlikely to best represent the quality of instruction that students were randomly assigned to.

The ITT estimates may meaningfully differ from observed teacher effectiveness, as measured by the FFT, when postrandomization sorting has taken place. To assess the extent of this difference, we next estimate Equation 2 by including the measure of *Teach* for the student's actual teacher for the 2010–2011 school year. Comparing the ITT estimates with observed teacher effectiveness will provide insight into the nature of student–teacher sorting and how this sorting may influence the relationship between teacher performance, as measured by the FFT, and student achievement. We discuss these findings in the Results section.

Accounting for Noncompliance With Randomization

Subsequent to the random assignment of classes of students to teachers, there was extensive noncompliance with teacher random assignment across each of the six district sites. In the MET study's full randomization sample, only 30% of students in Memphis complied with their initial teacher random assignment, whereas 54% of students who remained in the same school were noncompliers (22% of students moved to another teacher within the randomization block, while 32% of students remained in the school but ended up in a classroom outside of the exchange group randomization block; White & Rowan, 2012). Noncompliance likely reflects two types of post hoc sorting: (a) students requesting transfers away from their randomly assigned teacher's class and (b) teachers and/or principals purposefully matching students to teachers. If the movement of students within and outside of randomization blocks is nonrandom, then we can no longer assume that teacher instructional quality is orthogonal to student characteristics.

To better understand the patterns of noncompliance in our analytic samples, we first describe the characteristics of students and their actual teachers by student compliance status, with results summarized in Table 4. For both the ELA and math analytic samples, on average, students who complied with their random teacher assignment tended to be older, more likely to be ELL, more likely to receive free/reduced-price lunch, and less likely to be Black (but more likely to be Hispanic). While there are no differences in student achievement among compliers and noncompliers for students in the ELA analytic sample, students who remained with their random teacher assignment had lower math achievement in both 2010 and 2011 than the noncompliers. Moreover, while academically gifted students were more likely to comply among the ELA sample, they were less likely to do so among the math sample. Among students who did not remain with their initial random teacher assignment, they appear to end up with teachers who are less likely to be White, more likely to be Black, more likely to have a master's degree (or higher), and who have fewer years of experience teaching in the district.

We next explore the joint distribution of observed teacher performance with student compliance status to better understand whether compliance status is associated with observed teacher effectiveness. That is, were there systematic sorting patterns related to teacher effectiveness as measured by the FFT? We summarize these findings in Table 5. For each student's actual

TABLE 4
Summary Statistics by Student Compliance Status

	ELA sample		Math sample	
	Compliers	Noncompliers	Compliers	Noncompliers
Student characteristics				
Age	10.5*** (1.5)	9.8 (1.2)	10.2*** (1.4)	9.8 (1.2)
Fourth grade	25%***	32%	30%***	35%
Fifth grade	29%***	43%	31%***	39%
Sixth grade	17%***	11%	16%**	14%
Seventh grade	14%***	9%	11%***	5%
Eighth grade	15%***	5%	12%***	7%
Male	50%	49%	50%	48%
Gifted	12%***	9%	5%***	8%
Special education	6%	9%	7%*	8%
ELL	14%***	12%	17%***	11%
Free/reduced-price lunch	62%***	51%	65%***	49%
White	28%***	23%	25%***	28%
Black	27%***	35%	27%***	35%
Hispanic	35%***	30%	38%***	26%
Asian	8%***	10%	7%***	9%
Other race	3%	2%	3%	2%
Subject test score 2010	0.16 (0.92)	0.19 (0.96)	0.10*** (0.91)	0.21 (0.92)
Subject test score 2011	0.17 (0.91)	0.20 (0.92)	0.08*** (0.90)	0.21 (0.90)
Teacher characteristics				
Male	13%	12%	17%***	11%
White	65%***	61%	66%***	59%
Black	27%***	32%	24%***	35%
Hispanic	6%	6%	7%**	6%
Other race	1%	1%	3%***	0%
Years in district	7.8*** (6.8)	6.1 (4.4)	7.6*** (7.0)	6.6 (6.4)
Masters or higher	35%***	41%	34%***	47%
Students	6,951	1,829	5,903	2,031
Teachers	492	274	419	282
Schools	153	104	138	112

Note. Authors' calculations from MET data from the 2010–2011 school year. Mean (and standard deviation) reported. The ELA and math samples consist of all students with complete test score data from both the 2009–2010 and 2010–2011 school years. Teacher characteristics are for a student's actual subject-specific teacher during the 2010–2011 school year. Test scores have been standardized within district, subject, and grade. ELL = English language learner; MET = Measures of Effective Teaching. Differences between student compliers and noncompliers, by subject, statistically significant at the *10%, **5%, and ***1% levels.

2010–2011 teacher, we use the teacher's prerandomization (e.g., 2009–2010 school year) FFT scores.[7] Students who did not remain with their initial teacher random assignment were sorted to teachers with higher measures of *Teach*, both in ELA and in math. These results hold for almost all of the eight component-level measures as well as the domain-level aggregates for classroom environment and instruction.

Finally, to address the consequences of noncompliance on estimates of teacher effectiveness, we employ an instrumental variable (IV) strategy. Specifically, we use the randomly assigned teacher's previous year's measure of instructional

TABLE 5
FFT Descriptive Statistics by Student Compliance Status

	ELA			Math		
	All	Compliers	Noncompliers	All	Compliers	Noncompliers
Domain 2: *Classroom environment*	2.72 (0.33)	2.72*** (0.33)	2.75 (0.32)	2.68 (0.35)	2.67** (0.35)	2.69 (0.34)
2a: Creating an environment of respect and rapport	2.76 (0.37)	2.75*** (0.38)	2.79 (0.35)	2.69 (0.39)	2.69 (0.39)	2.69 (0.40)
2b: Establishing a culture for learning	2.57 (0.39)	2.56*** (0.39)	2.62 (0.38)	2.54 (0.38)	2.53*** (0.39)	2.58 (0.35)
2c: Managing classroom procedures	2.73 (0.37)	2.73 (0.37)	2.74 (0.36)	2.69 (0.39)	2.69** (0.39)	2.71 (0.40)
2d: Managing student behavior	2.84 (0.37)	2.83*** (0.37)	2.86 (0.37)	2.77 (0.39)	2.76*** (0.39)	2.79 (0.38)
Domain 3: *Instruction*	2.45 (0.32)	2.44*** (0.32)	2.47 (.31)	2.41 (.32)	2.40*** (.32)	2.43 (.33)
3a: Communicating with students	2.68 (0.34)	2.68 (0.34)	2.69 (0.34)	2.62 (0.35)	2.62*** (0.35)	2.65 (0.35)
3b: Using questioning and discussion techniques	2.31 (0.39)	2.30*** (0.39)	2.34 (0.39)	2.18 (0.37)	2.17*** (0.37)	2.21 (0.39)
3c: Engaging students in learning	2.49 (0.37)	2.48*** (0.37)	2.53 (0.36)	2.46 (0.39)	2.44*** (0.40)	2.49 (0.36)
3d: Using Assessment in Instruction	2.29 (0.38)	2.29*** (0.39)	2.32 (0.34)	2.36 (0.39)	2.36 (0.38)	2.35 (0.42)
Teach	2.58 (0.30)	2.58*** (0.29)	2.61 (0.29)	2.54 (0.32)	2.53*** (0.32)	2.56 (0.31)
Students	8,750	6,951	1,799	7,889	5,903	1,986
Teachers	578	492	271	521	419	275
Schools	163	153	103	156	138	110

Note. Means reported (with standard deviation in parentheses) are the 2009–2010 FFT scores for students' actual 2010–2011 teachers. The value for Domain 2 is an average of items 2a–2d; the value for Domain 3 is an average of items 3a–3d. *Teach* is an average of the eight items (2a–3d) across the two domains. Each item is rated on an integer scale of 1 to 4, where a score of 1 is unsatisfactory, 2 is basic, 3 is proficient, and 4 is distinguished. For the ELA sample, 30 students (of the 8,780 students in the ELA analytic sample) sorted to teachers without a 2010 FFT score for ELA instruction. For the math sample, 45 students (of the 7,934 students in the math analytic sample) sorted to teachers without a 2010 FFT score for math instruction. Differences between compliers and noncompliers, by subject, statistically significant at the *10%, **5%, and ***1% levels. FFT = Framework for Teaching.

practice ($Teach_{ik,t-1}$) as an instrument for the observed instructional practice ($Teach_{ijt}$) of a student's teacher during the 2010–2011 school year. As a student was randomly assigned to teacher k within a randomization block, the unobservable component (ε_{ijt}) of student i that is correlated with their achievement ($Achieve_{ijt}$) should be independent of the observed prior-year teaching practice of their randomly assigned teacher.[8] The following two-stage system summarizes the IV approach:

$$Teach_{ijt} = \alpha_0 + \alpha_1\left(Teach_{ik,t-1}\right) + \alpha_2\left(Achieve_{i,t-1}\right) + \mathbf{X}_{it}\Gamma + \mathbf{Z}_{jt}\varphi + \theta_r + \varepsilon_{ijt}. \quad (3)$$

$$Achieve_{ijt} = \beta_0 + \beta_1\left(\widehat{Teach_{ijt}}\right) + \beta_2\left(Achieve_{i,t-1}\right) + \mathbf{X}_{it}\Gamma + \mathbf{Z}_{jt}\varphi + \theta_r + \varepsilon_{ijt}. \quad (4)$$

In Equation 3, $j = k$ if student i complied with her initial teacher random assignment, and all other variables are defined as in Equation 2. The IV estimate (b_{IV}) is therefore $cov(Achieve_{ijt}, Teach_{ik,t-1}) / cov(Teach_{ijt}, Teach_{ik,t-1})$, conditional on the randomization block. Under full compliance (e.g., $j = k\ \forall\ i$ within randomization block θ_r), using the previous year's practice measure as an instrument allows us to remove any bias due to measurement error, generating an unbiased estimate of a teacher's true effect on student achievement.[9] With

TABLE 6
Intent to Treat (ITT) Estimates

	ELA			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Teach	0.011 (0.039)	−0.001 (0.037)	0.008 (0.037)	0.049 (0.036)	0.053 (0.035)	0.061* (0.033)
Student characteristics		x	x		x	x
Teacher characteristics			x			x
R^2	.6341	.6468	.6477	.7034	.7106	.7117
Students	8,780	8,780	8,780	7,934	7,934	7,934
Randomized teachers	593	593	593	537	537	537
Randomization blocks	259	259	259	246	246	246

Note. Each column represents a separate regression. Coefficients are in standard deviation units. Student covariates are age, race, gender, ELL, FRPL, special education status, and gifted status. Teacher covariates are gender, race, degree status (masters+), and experience in district. All regressions include controls for randomization block fixed effects and controls for prior (math or ELA) student achievement (2010). Robust standard errors (clustered at the randomized teacher level) reported in parentheses. *Teach* is the mean of the eight FFT items from the 2009–2010 school year for students' randomly assigned teachers for the 2010–2011 school year. ELL = English language learner; FRPL = Free/reduced-price lunch; FFT = Framework for Teaching.
Coefficients are statistically significant at the *10%, **5%, and ***1% levels.

noncompliance, the denominator of the IV estimate [$cov(Teach_{ijt}, Teach_{ik,t-1})$] will be smaller relative to its magnitude under full compliance; as a result, the IV estimate will be scaled up, representing the treatment-on-treated effect of teacher effectiveness. We discuss these results below.

Results

We first consider the intended effect of the teacher a student was randomly assigned to on that student's performance on state achievement tests. These ITT results are summarized in Table 6. We find that there is no effect, either in magnitude or in statistical significance, of the randomly assigned teacher, as measured by their 2010 ELA performance on the FFT (e.g., *Teach*). These results are consistent even when we control for student and teacher characteristics. In contrast, we find modest, marginally significant effects of the randomly assigned teacher, as measured by 2010 math *Teach*, on student math achievement. Specifically, we see an improvement in student achievement of approximately 0.05 to 0.06 standard deviations when assigned to a teacher who is proficient rather than basic on the FFT. This estimate is robust in magnitude, but not in statistical significance, across specifications that include student and teacher characteristics.

This ITT estimate is often the policy-relevant parameter of interest when considering the potential impact of a new policy or educational intervention, under the assumption that individuals may or may not comply with their initial assignment.[10] In this context, we consider the ITT estimate to be a lower bound on teacher effectiveness, as measured by the FFT. However, this estimate does not provide any insight into the observed relationship between teacher effectiveness and student performance when students and teachers are purposefully matched. To assess the observed relationship, we next investigate the relationship between the quality of the teacher (as measured by the FFT) to whom a student actually received instruction from and that student's academic performance. These results are summarized in Table 7. We find that, on average, higher performing teachers both in ELA and in math, as measured by the FFT, are associated with higher student performance. For ELA, the difference in measured performance, from basic to proficient, is associated with reading improvement gains on the order

TABLE 7
Observed Teacher Effectiveness: Relationship Between Teaching Practice and Student Achievement

	ELA			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Teach	.121*** (0.040)	.107*** (0.039)	.109*** (0.038)	.141*** (0.042)	.135*** (0.042)	.161*** (0.039)
Student characteristics		x	x		x	x
Teacher characteristics			x			x
R^2	.635	.647	.647	.704	.711	.713
Students	8,780	8,780	8,780	7,934	7,934	7,934
Actual teachers	581	581	581	528	528	528
Randomization blocks	259	259	259	246	246	246

Note. Each column represents a separate regression. Coefficients are in standard deviation units. Robust standard errors clustered at the section level are reported in parentheses. All regressions control for randomization block fixed effects and a student's prior (2010) ELA or math achievement. *Teach* is the mean of the eight FFT items for a student's actual 2011 teacher. Student characteristics include age, race, gender, ELL status, free/reduced-price lunch status, special education status, and gifted status. Teacher characteristics include gender, race, educational attainment, and years of experience in the district. Classroom characteristics include the number of students in the class section, percent of students who are male, special education, ELL, free/reduced-price lunch, gifted, the percent of students by race, average age, and average 2010 ELA or math achievement. FFT = Framework for Teaching; ELL = English language learner.
Coefficients are statistically significant at the *10%, **5%, and ***1% levels.

of 0.11 standard deviations. These associations are robust and highly statistically significant across model specifications. For math, the observed relationship between teacher effectiveness and student performance is on the order of 0.14 to 0.16 standard deviations; again, these estimates are highly significant and robust to the inclusion of student and teacher characteristics.[11]

As previously discussed, extensive and systematic sorting occurred among students postrandomization. While students appear to be sorting to higher performing teachers, as measured by the FFT, we are unable to determine the direction of the potential bias induced by this sorting. Specifically, we are unable to empirically assess whether unobserved characteristics of students and teachers are systematically related in ways that affect both teacher practice and student achievement. Therefore, given the potential (and likely) bias present in estimates of teacher performance due to student noncompliance with the initial randomization, we employ an IV strategy to identify teacher effectiveness using multiple classroom observation scores on the FFT. These IV results are summarized in Table 8.

The first-stage results provide an estimate of the relationship between the randomly assigned teacher's measure of *Teach* for the 2009–2010 school year and the student's actual teacher's measure of *Teach* for the 2010–2011 school year. Note that the coefficient on the first-stage estimate would equal 1 if all of the following conditions were satisfied: (a) Students perfectly complied with their random teacher assignment; (b) the 2010 *Teach* measure was perfectly predictive of a teacher's 2011 *Teach* measure; and (c) there was no systematic measurement error arising from, for example, differences among raters assigned to evaluate different lessons (from either the same or different school years). Conversely, if no students complied with their random teacher assignment and the 2010 *Teach* measure for the randomized teacher was completely uncorrelated with the actual teacher's 2011 *Teach* measure, then the first-stage estimate would equal 0. We find that, for both the ELA and math analytic samples, the average association between the 2010 and 2011 measures of *Teach* is between 0.13 and 0.14. This relationship is highly significant and consistent across

TABLE 8
Instrumental Variable Estimates

	ELA			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Teach	0.076 (0.281)	−0.009 (0.271)	0.003 (0.277)	0.395 (0.295)	0.419 (0.293)	0.254 (0.237)
Student characteristics		x	x		x	x
Teacher characteristics			x			x
I.V. (*F* statistic)	.142*** (*F* = 9.43)	.142*** (*F* = 9.44)	.136*** (*F* = 8.35)	.126*** (*F* = 8.03)	.126*** (*F* = 8.07)	.132*** (*F* = 10.37)
R^2 (second Stage)	.635	.647	.647	.702	.708	.713
Students	8,780	8,780	8,780	7,934	7,934	7,934
Actual teachers	581	581	581	528	528	528
Randomization blocks	259	259	259	246	246	246

Note. Each column represents a separate regression. Coefficients are in standard deviation units. Student characteristics include age, race, gender, ELL, FRPL, special education, and gifted status. Teacher characteristics include gender, race, degree status (masters+), and experience in district. All regressions include controls for randomization block fixed effects and controls for prior (math or ELA) achievement (2010). Robust standard errors (clustered at the section level) reported in parentheses. *Teach* estimates the impact of the 2011 FFT score of the student's actual teacher when instrumenting with the 2010 FFT score of the student's randomized teacher. ELL = English language learner; FRPL = Free/reduced-price lunch; FFT = Framework for Teaching. Coefficients are statistically significant at the *10%, **5%, and ***1% levels.

specifications. While these results are stable, the magnitude of the year-to-year relationship of a teacher's FFT score is modest. We note that for students' actual ELA teachers for the 2010–2011 school year, the correlation between their 2010 and 2011 measures of *Teach* is .526. For students' actual math teachers for the 2010–2011 school year, the correlation between their 2010 and 2011 measures of *Teach* is .521. Given these modest year-to-year, within teacher correlations (coupled with the extent of student noncompliance), it is not surprising that the first-stage results yield the reported magnitudes, while also constraining our ability to identify teacher effectiveness. Indeed, as we look to our second-stage results, the IV estimates for the relationship between *Teach* and student achievement do not approach statistical significance. We discuss the policy implications of these findings in the next section.

In light of the extent of noncompliance with randomization, and the limited ability to estimate teacher effectiveness using 2 years of FFT scores, the ITT and observed estimates provide an empirical range for understanding the relationship between teacher effectiveness, as measured by the FFT, and student achievement. One useful way to contextualize these estimated ranges is to compare them with a relevant benchmark. Hill, Bloom, Black, and Lipsey (2008) provide such a benchmark using nationally normed standardized tests; the authors find that the average expected annual growth in Grades 4 to 8 in reading and math is 0.30 and 0.41 standard deviations, respectively. On the basis of this benchmark, we first consider the ITT effect. In particular, the expected difference in math achievement between students randomly assigned to proficient teachers and students randomly assigned to basic teachers (based on their 2009–2010 FFT ratings) would be 12% to 15% of an average year of learning by students nationwide—that is, 1.2 to 1.5 months of learning in a 10-month school year. We next consider the observed relationship, postrandomization, between actual teacher performance and student achievement. The difference in reading achievement between students who attended classes

(during the 2010–2011 school year) with proficient teachers and students of basic teachers was 36% of an average year of reading learning by students nationwide—that is, 3.6 months of learning in a 10-month school year. For math, the actual difference in achievement was 34% or 3.4 months of learning.

Discussion

To improve teacher quality, recent education policy has emphasized empirically based evaluation systems using multiple measures of teaching performance. Among these measures, observational assessments of a teacher's instructional practice have emerged as a critically important component. As districts and states respond to policy efforts and incorporate observational protocols like the FFT into their evolving teacher evaluation systems, we need to more fully understand whether these measures are able to reliably identify teacher effectiveness. Understanding this relationship is particularly salient for policy purposes, given the overwhelming influence of observational measures in determining a teacher's summative, year-end evaluation scores.

We find that information about teacher effectiveness gathered across FFT component scores can be usefully aggregated into a simple average. This straightforward approach provides practical guidance for principals and administrators when using the FFT to measure teacher effectiveness. Moreover, these average scores, whether generated from ELA or math lessons, are highly correlated with student performance. On average, student achievement is higher among teachers who receive higher FFT ratings. However, for policy to compel educational administrators to make high-stakes personnel decisions, such as teacher tenure, by relying heavily on FFT measures ignores one of the key drivers of this relationship—the systematic sorting of students to teachers. We find consistent patterns of noncompliance with randomization that moves students to teachers with higher FFT scores. Such nonrandom sorting limits the ability of teacher performance measures to provide a valid estimate of a teacher's contribution to student learning, thereby constraining policymakers' and school leaders' ability to identify truly effective teachers.

In interpreting these results, there is an important distinction to be made between identifying the effect of classroom observation scores (such as those produced by the FFT or other such observation protocols) on student achievement, as opposed to identifying teacher effectiveness using the FFT as a measure of teacher performance. The design of the MET study enables the latter, and not the former. The random assignment of teachers to students enables an unbiased estimate of the average causal effect of teachers on student achievement, and here we reflect many current teacher evaluation systems by using FFT scores as our measure of effectiveness. In contrast, the design does not allow for an estimate of the causal effect of manipulating a teacher's FFT rating on student achievement. The observational score is thus a marker for teacher effectiveness, but not an intervention.

Implicit in this distinction is the impossibility of either fully capturing or randomly assigning instructional quality. While better teachers, on average, may receive higher FFT ratings, there are likely other aspects of teacher quality that are salient to student learning but not measured by the FFT. As an example, suppose our goal was to identify the effects of FFT scores themselves on student achievement. To do so, imagine we randomly assign teachers to different instructional regimes that are aligned with different levels of the FFT. Even under these experimental conditions, it would be impossible to demonstrate that any associated changes in student achievement would be attributable to the randomly assigned FFT level and not other shifts in instruction or teacher traits, unless we believed that the FFT measured all aspects of a teacher that influence student achievement, and all of those aspects perfectly aligned with the randomly assigned instructional regimes.

Disentangling the effect of quality instruction on student achievement is further complicated by the fact that instruction undoubtedly interacts with numerous factors, including the composition of students in a given class. The mix of students on observed and unobserved dimensions will vary both across teachers within a school, as well across classes taught by the same teacher. As the quality of instruction is influenced by the particular mix of students a teacher works with in a given class, observed measures of a teacher's classroom

instructional practice, such as those provided by the FFT, are very much context-specific.

Given that instructional quality is shaped by the composition of students in a teacher's class, and in light of the noncompliance with randomization in our data, our most policy-relevant findings are contained within the ITT and observed estimates—the expected and observed influences of teacher quality on student achievement. The ITT estimates suggest that expected teacher effectiveness, as measured by FFT scores, is context-specific across years, as the randomly assigned teacher's measure of effectiveness during the 2009–2010 school year is only weakly associated with student achievement in the subsequent school year. Indeed, for our sample of fourth- through eighth-grade teachers, we find that the year-to-year correlation in observed effectiveness is .53 for reading and .52 for math, as measured by *Teach*. So, while critics of value-added measures point to the year-to-year variability as a limitation of value-added scores in identifying highly effective teachers, our results suggest a similar limitation for observation-based measures of teacher practice, based on the FFT.[12]

Our evidence also suggests fairly extensive, post hoc, nonrandom sorting of students to teachers. While this type of sorting limited our ability to causally identify teacher effectiveness with 2 years of classroom observation scores, it also likely reflects the intentional and potentially beneficial matching of students to teachers. Given this assortative matching process, the observed estimates approximate an upper range on the association between observed instructional practice and student achievement. We would note that despite the extent of observed noncompliance with randomization, assortative matching was likely more limited than under a natural context, and thus the optimizing of teacher–student matching not fully realized. If this is the case, then the observed estimates may understate the influence of teacher quality on student achievement.

It is also important to recall that this study considers the status quo relationship between teacher effectiveness and student achievement, without incorporating the FFT protocol's intended feedback sessions and opportunities to triangulate with professional development. The FFT system was constructed specifically to enable educator growth, and therefore this study does not speak to either the potential impacts on student and teacher performance when the FFT protocol is fully implemented, or how performance can be shaped over time (see Steinberg & Sartain, in press; Taylor & Tyler, 2012). Indeed, this potential for professional development embedded within the complete FFT protocol is one of the compelling reasons for its use, as compared with value-added scores, which provide little guidance for teachers on how to improve their practice. This study does not speak to the potential benefits of structuring these growth opportunities into teacher evaluation systems, and further investigation into this area is warranted. The limitations of the FFT as a measure of teacher effectiveness alone may also be overcome by other developing performance measures. For example, while practical limitations on using value-added scores across all teachers will remain, efforts by states and districts to develop SLOs for all grades and subjects may prove useful for use in conjunction with classroom observation measures, for both accountability and development purposes. As the field continues to respond to policy shifts, it will be important for research to support these changing practices.

Ultimately, our results lead us to question the wisdom of solely relying on observational protocols like the FFT for the purposes of evaluating teachers in a formal system that will affect decisions relating to tenure, performance pay, and other key personnel decisions. As new teacher evaluation systems are being developed, more districts are paying careful attention to teacher performance measures based on student test scores, but they may not be giving enough thought to the role of observational measures in those same evaluation systems. For the majority of teachers in untested subjects and grades, observation measures such as the FFT will likely remain a key component of teacher evaluation. Our results indicate that equal care needs to be taken when using FFT measures for high-stakes purposes; the findings also highlight the need for stakeholders to understand the limitations of the FFT. Decisions made about teacher effectiveness, either to praise or censure, should not rely solely on measures from the FFT until we have a

better understanding of how it captures teacher performance over time.

Authors' Note

Authors are listed in alphabetical order.

Acknowledgments

We thank Guanglei Hong and Jonah Deutsch for valuable feedback on early stages of this work. We also benefitted from helpful suggestions from three anonymous reviewers and the journal editor, and conference participants at the National Academy of Education Annual Meeting, Society for Research on Educational Effectiveness, and American Educational Research Association.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding from the Measures of Effective Teaching (MET) Early Career Grantee program is gratefully acknowledged.

Notes

1. For example, the Pennsylvania Department of Education, a recipient of a Race to the Top (RTTT) grant, is currently implementing its Educator Effectiveness plan for evaluating teacher performance based on four components—classroom observation of teacher instruction (50%), building-level data, such as the school's graduation rate and promotion rate (15%), district-determined SLOs (20%), and value-added estimates based on student test score data (15%).

2. Steinberg and Sartain (in press) investigate how the implementation of the Framework for Teaching (FFT) protocol, including teacher–principal conferences, affects student achievement.

3. Please see the online appendix, available at http://epa.sagepub.com/supplemental for a detailed description of the FFT domains and components used in the Measures of Effective Teaching (MET) study.

4. On average, each teacher was rated on four subject-specific videos (i.e., an average of four videos for reading instruction or four videos for math instruction), with video recordings spread over time to capture greater representativeness of teacher instruction. See White and Rowan (2012) for a detailed discussion of the process of video recording and scoring teachers' lessons.

5. The only exception for the student characteristics was for the free/reduced-price lunch indicator, which was 24.9% missing in the ELA sample and 22.7% missing in the math sample.

6. Empirical work conducted elsewhere using the FFT also aggregated classroom observation scores by averaging the components of the FFT across multiple raters (Kane, Taylor, Tyler, & Wooten, 2011), and MET researchers aggregated classroom observation scores from the FFT (by averaging across the underlying components), and incorporated this aggregated FFT score into a composite measure of teacher performance (which included value-added scores, student surveys, and classroom observation scores from the Danielson FFT; Kane, McCaffrey, Miller, & Staiger, 2013). To justify the aggregation of FFT scores across raters and observations within a teacher's section, Kane et al. (2013) refer to Mihaly, McCaffrey, Staiger, and Lockwood (2013), who note that "FFT has four component measures used to assess each of these two domains and we use the average of the eight component scores. The section-level score was calculated by taking the average of the ratings from multiple raters and multiple video recordings of the section," and the authors indicate that "the combined measure is justified for observations since preliminary analysis suggested one factor and we expect districts to combine domains into a single measure" (Mihaly et al., 2013, p. 12). Our work herein also revealed one underlying factor of effective teaching using the FFT.

7. We use the teacher's prerandomization FFT scores to avoid any contamination that may arise in measuring the postrandomization FFT scores due to higher achieving students sorting to higher performing teachers. If such positive sorting is occurring, the interaction of higher achieving students with higher achieving teachers may artificially inflate the teacher's effectiveness measure. The patterns we observe using the prerandomization FFT scores hold when we look at the distribution of student compliance status and teachers' postrandomization FFT scores (these results are available upon request).

8. This is the exclusion restriction necessary for the prior-year practice measure of the randomly assigned teacher to be a valid instrument, and may be written as $cov(\varepsilon_{ijt}, Teach_{ik,t-1}) = 0$.

9. To see this, let $R1$ be the randomly assigned teacher's prior-year practice measure; $R2$ is the randomly assigned teacher's current year practice measure; $A2$ is the actual teacher's current year practice measure; and Y is a student achievement outcome. Suppose the true score of the randomly assigned teacher's practice, T, is time invariant and is measured with random error: $R1 = T + e1$ and $R2 = T + e2$. If the

outcome equation of interest is $Y = B_0 + B_1(A2) + u$, where u is a random error term, the IV estimate (b_{IV}) = $\frac{cov(Y,R1)}{cov(A2,R1)} = \frac{cov(Y,T+e1)}{cov(A2,T+e1)} = \frac{cov(Y,T+e1)}{cov(R2,T+e1)} = \frac{cov(Y,T+e1)}{cov(T+e2,T+e1)} = \frac{cov(Y,T)}{var(T)}$, representing the effect of true teacher practice on student achievement.

10. One can also think of the intent-to-treat (ITT) effect as the impact of the *offer* of a teacher's instructional practice on student achievement. As an analogy, suppose we randomly assigned students to receive vouchers for free afterschool tutoring, and some students used the vouchers while others declined the offer. The ITT effect would estimate the impact of the offer of free afterschool tutoring, and not the effect of actually receiving tutoring. The impact of receiving tutoring, like the impact of receiving a teacher's instructional practice, is considered the treatment-on-the-treated (TT) effect, which we consider in the context of the IV estimation.

11. In analyses not presented here, we explored both the ITT and observed relationship between the individual FFT components and student achievement. The patterns were confirmatory with our analyses using the composite *Teach* measure. In particular, we find a positive and significant relationship for the observed estimates using the actual teacher's 2011 component measures and no meaningful or significant patterns for the ITT estimates, using the randomly assigned teacher's 2010 component measures. As we do not find meaningful differences from our main analyses, we have elected to omit these models from the article for the sake of brevity. However, the findings are available upon request.

12. In the case of value-added estimates of teacher effectiveness, evidence suggests that the year-to-year correlation in value-added measures, across multiple studies, is on the order of .2 to .6 (Glazerman et al., 2010). McCaffrey, Sass, Lockwood, and Mihaly (2009) find that, for elementary and middle school teachers, the year-to-year correlations are .2 to .5 and .3 to .7, respectively, whereas Goldhaber and Hansen (2010) find that, for a sample of fifth-grade teachers in North Carolina, the year-to-year correlations in reading and math achievement are .3 to .4 and .5 to .6, respectively.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the midwest region* (Issues & Answers Report, REL 2007–No. 030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, REL Midwest.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, *79*, 79–107.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

Goldhaber, D. D. (2002). The mystery of good teaching. *Education Next*, *2*, 50–55.

Goldhaber, D. D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance* (CEDR Working Paper #2010-3). Seattle: University of Washington.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*, 103–123.

Kane, T. J., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Retrieved from MET Project website: http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved from MET Project website: http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*, 587–613.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship

between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54–78.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*, 572–606.

Mihaly, K., McCaffrey, D. F., Staiger, D., & Lockwood, J. R. (2013, January 8). *A composite estimator of effective teaching*. Retrieved from MET Project website: http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53.

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *The Elementary School Journal*, *88*, 166–187.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*, 537–571.

Steinberg, M. P., & Sartain, L. (in press). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*, 3628–3651.

Watson, J. G., Kraemer, S. B., & Thorn, C. A. (2009). *The other 69 percent*. Washington, DC: Center for Educator Compensation Reform, Office of Elementary and Secondary Education, U.S. Department of Education.

White, M., & Rowan, B. (2012). *A user guide to the "core study" data files available to MET early career grantees*. Ann Arbor: Inter-University Consortium for Political and Social Research, The University of Michigan.

Authors

RACHEL GARRETT is a researcher at the American Institutes for Research. Her research interests include educator quality, policy and program evaluation, English Language Learners, and math learning.

MATTHEW P. STEINBERG is an assistant professor in the Graduate School of Education at the University of Pennsylvania. His research focuses on teacher evaluation and human capital, urban school reform, and school climate and safety.

Manuscript received October 10, 2013
First revision received January 4, 2014
Second revision received April 1, 2014
Third revision received April 30, 2014
Accepted May 2, 2014