Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools

Simon Burgess, *University of Bristol*

Shenila Rawal, *Oxford Partnership for Education Research and Analysis*

Eric S. Taylor, *Harvard University*

This paper reports on a field experiment in 82 high schools trialing a low-cost intervention in schools' operations: teachers working in the same school observed and scored each other's teaching. Students in treatment schools scored 0.07 student standard deviations higher on math and English exams. Teachers were further randomly assigned to roles—observer and observee—and students of both types benefited, observers' students perhaps more so. Doubling the number of observations produced no difference in student outcomes. Treatment effects were larger for otherwise low-performing teachers.

I. Introduction

Performance evaluation is ubiquitous in modern economies. Employers observe and try to measure the job performance of their employees with the

We first thank the Education Endowment Foundation for generous financial support, the National Foundation for Educational Research for its input, and the

Submitted November 5, 2019; Accepted November 30, 2020; Electronically published August 2, 2021.

Journal of Labor Economics, volume 39, number 4, October 2021.
 https://doi.org/10.1086/712997

goal of improving that performance. In typical practice, such performance measures are combined with incentives; the measures determine bonuses, dismissals, promotions, and so on. Yet evaluation might improve performance even without attaching (explicit) incentives. The process of being measured itself could reveal new information about an individual's current skills or effort, or it could emphasize the employer's expectations and thus motivate or direct an individual's efforts to improve. Moreover, typically the sole focus of the process is the performance of the employee being observed. The role of the observer is rarely considered, either taken as simply another task for management or outsourced to external experts. Yet in some professions and for some tasks, the observer may also gain from the process, picking up new ideas and practices.

In this paper's experiment we focus on low-stakes peer evaluation among teachers. We estimate the effects of evaluation, if any, where the potential mechanisms for those effects cannot rely on explicit incentives but where effects can arise through learning more about own and coworker performance. We set up and report on an experiment in which teachers assessed and scored each other's job performance, using structured observations of teaching in the classroom, and discussed the results together. These peer evaluations were "low stakes" in the sense that there were no formal incentives or consequences attached to the scores, although there may have been informal incentives like social pressure or career concerns.

A distinctive new contribution of our experimental design is that we can separately estimate effects on the teachers being evaluated (observees) and teachers serving as peer evaluators (observers). Within each of the treatment schools, individual teachers were randomly assigned to be an observee, an observer, or both. Observers' performance may suffer, for example, because they reallocate effort to conducting evaluations; such potential losses are an important opportunity cost of peer evaluation to weigh against benefits for observees. Observers' performance may also benefit from the opportunity to learn from colleagues and to reflect on their own teaching practice. We believe that our evidence is the first to isolate the impacts on observees and observers experimentally.

Department for Education for access to the National Pupil Database. We are indebted to Julia Carey and her team for outstanding project management and to the schools and teachers for their participation. Thanks also to Anna Vignoles, Ellen Greaves, Hans Sievertsen, and seminar participants at the INSEAD/Field Days Workshop, the Institute of Labor Economics (IZA) Economics of Education Workshop, the University of Sydney, the Association for Education Finance and Policy (AEFP), and the Association for Public Policy Analysis and Management (APPAM), who provided helpful comments and questions. Contact the corresponding author, Eric S. Taylor, at eric_taylor@harvard.edu. Information concerning access to the data used in this paper is available as supplemental material online.

A second new contribution is to examine the intensive margin of the number of peer observations. In a random half of treatment schools, math department teachers were expected to be observed twice as many times as their English department colleagues; in the other half, the English department was assigned to the double dose. Ours is the first experimentally induced variation of which we are aware.

The effect of evaluation on performance is of growing interest to school administrators and policy makers tasked with managing the teacher workforce. Econometric research, beginning in the 1970s and accelerating in the past decade, demonstrates educationally and economically significant differences between teachers in their contributions to student learning. However, we still understand comparatively little about how to select better teachers or how to train teachers before or on the job (for a review, see Jackson, Rockoff, and Staiger 2014). The importance of teachers, but the lack of management tools, has prompted new attention to teacher performance evaluation in recent years. One common proposal is probationary screening: measure on-the-job performance early and dismiss observed low performers (Gordon, Kane, and Staiger 2006). While the first steps are intuitive, the equilibrium effect of this proposal likely depends more on how labor supply responds than on the measurement process (Staiger and Rockoff 2010; Rothstein 2015). A second common proposal is that the process of evaluation itself should, if done well, improve performance (Milanowski and Henemen 2001). This proposal is made largely without empirical evidence, although we discuss notable exceptions below, and the present experiment is designed in part to help close that evidence gap.[1]

Our peer evaluation experiment took place over two school years in 82 secondary schools in England. In treatment schools, year 10 and 11 (age 15 and 16) math and English teachers were asked to participate in a new program of peer classroom observations, with observations scored using a structured, well-established rubric. Control schools continued business as usual, which generally did not involve teacher classroom observations by peers. The main analysis sample includes just over 28,000 students and approximately 1,300 teachers. The outcome is students' math and English scores on the General Certificate of Secondary Education (GCSE) exam.

The paper details three main results. First, the program of low-stakes teacher peer evaluation meaningfully improves student achievement in math and English. Students in treatment schools scored 0.073 student standard deviations (σ) higher, on average, than their counterparts in control schools.

[1] Evaluative measures are also an important input to the long-standing proposals for teacher pay for performance. For a review of the theory and empirical literature, see Neal (2011) and Jackson, Rockoff, and Staiger (2014), along with a notable more recent example, Dee and Wyckoff (2015).

Second, the program benefits do not depend on the number of peer observations, at least over the range induced by the experiment. The difference between outcomes in high- and low-dose conditions was -0.002σ, even though high-dose departments completed nearly twice as many observations (2.9 vs. 1.6 per observee). Third, student achievement improved in the classes of both observee and observer teachers. Improvements for observers rule out the concern that benefits for observees' students might come at the expense of losses for observers' students, but those improvements also raise new questions about potential mechanisms, which we discuss in section V. We cannot reject the null hypothesis that observers and observees benefited equality, but the point estimates consistently favor observers.

Most of the potential mechanism hypotheses are more detailed than we can test empirically; however, we do test whether treatment effects differ across quantiles of the teacher performance distribution. The quantile treatment effects (QTEs) are broadly consistent with the hypothesis that teachers learned new skills. Lower-performing teachers improved more under treatment than did higher-performing teachers, although these results are relatively imprecisely estimated.

Our paper contributes most directly to the literature on how evaluation affects teacher performance. Most of that literature focuses on the effect of incentives; the (quasi)-experimental comparison is teachers who face different monetary bonuses or the dismissal threats (for reviews, see Neal 2011; Jackson, Rockoff, and Staiger 2014). Our first contribution is to focus on the effect of evaluation measures; our experimental comparison is measures versus no measures, with no (explicit) incentives attached. Existing evidence on the effect of measures is quite scarce. Close to our work, Rockoff et al. (2012) study the effect of providing teacher value-added scores to schools—scores that measure a teacher's output based on her contribution to student test scores. In contrast to the input-focused scoring rubric in our experiment, output-focused value-added scores contain no suggestions for how to improve. While Rockoff et al. (2012) find small improvements in performance, an open question is whether the input-focused classroom observation measures might generate larger improvements, as their advocates have suggested (Milanowski and Henemen 2001).

Second, compared with prior work, this experiment more sharply identifies the effect of classroom observation separate from incentives. Three other (quasi) experiments, Taylor and Tyler (2012), Steinberg and Sartain (2015), and Briole and Maurin (2019), also find that teacher performance improves, as measured by student tests, when the teacher is evaluated by classroom observation. However, in all three cases the evaluation program studied was the formal evaluation program run by the school system, with explicit incentives attached to the scores. It is unclear how important the explicit incentives were in generating the changes those papers document. For example, in the Taylor and Tyler (2012) case the main stated incentive

was the threat of dismissal for repeated low scores, but in practice few teachers were ever actually dismissed. In the Briole and Maurin (2019) case scores partly determined wages. In the current experiment the program was organized external to the school system with no explicit incentives.

Last, our experiment contributes novel evidence on a practical management question: how to design "mentor-mentee" or "advisor-advisee" relationships for teachers. This paper is the first of which we are aware to randomly assign teacher roles in such relationships. Nearly all existing evidence comes from settings where the "advisor" or "evaluator" is (i) a formal job, with training, filled by an experienced high-performing teacher; or (ii) a school administrator.[2] Two notable exceptions, Papay et al. (2020) and Murphy, Weinhardt, and Wyness (2021), are cases where, as in our experiment, the advisor-advisee relationships are between peer coworkers. Both interventions also involved teachers watching each other teach, although in neither case were teachers scored during observations. However, in both cases the benefits to teacher performance apparently arose from intentionally matching strong teachers to weak teachers. This strong-to-weak feature is often assumed necessary for successful advisor-advisee relationships. Our results question the conventional wisdom that observers and mentors must be selected for a history of high performance, and the benefits to observers suggest a sizable missed opportunity in the design of such programs.

II. Treatment, Setting, and Data

This paper reports on a new field experiment in which coworker teachers observed each other teaching in the classroom, scored performance using a detailed rubric, and discussed the results. The intervention was conducted in secondary schools in England, focusing on years 10 and 11 math and English teachers, over two school years, 2014–15 and 2015–16. This section describes the treatment in detail, the study design, data, and other key features of the setting. Additional details of the experiment are provided in appendix B (apps. A–D are available online).

A. Random Assignment Design and Covariate Balance Tests

1. Schools and Departments

The experiment involved randomizing aspects of the intervention at three levels: school, department, and teacher. We first randomly assigned

[2] In the case of teacher evaluation specifically, examples of type i include Taylor and Tyler (2012), Dee and Wyckoff (2015), and Briole and Maurin (2019), and examples of type ii include Steinberg and Sartain (2015) and Dee and Wyckoff (2015). For a recent review of the broader advisor-advisee or coaching literature, see Kraft, Blazar, and Hogan (2018).

82 schools, half to the treatment—the new peer observation program—and half to a business-as-usual control condition.[3] We describe the recruitment and characteristics of the 82 schools below, as well as what "business as usual" means for these schools. Schools were assigned within eight randomization blocks defined by the interaction of three indicator variables, with each indicator equaling 1 if the school was above the sample median for (i) percentage of students who are white, (ii) percentage of students eligible for free school meals, and (iii) prior student achievement growth at the school (school value-added scores).

Table 1 shows the conventional pretreatment covariate balance tests relevant to judging the success of random assignment. Column 2 shows the test for random assignment of schools to treatment and control. Schools were well balanced on observables. None of the differences are statistically significant at any conventional levels, except the Income Deprivation Affecting Children Index (IDACI) score ($p = .065$). A given neighborhood's IDACI score is the proportion of children under 16 living in a low-income household; a student's IDACI value is the score for the neighborhood where they live.

Second, we randomly assigned departments to either a "high-dose" or a "low-dose" condition. Half of the treatment schools were randomly assigned to the condition: high dose for the math department, and low dose for the English department. The other half of treatment schools took the opposite: low math, high English. In the low-dose departments observee teachers were expected to be observed 6 times per year; high dose doubled the ask to 12. This random assignment of schools to dose conditions was independent of the main treatment-control random assignment; however, the dose randomization was within the same eight blocks.

Column 3 of table 1 shows the covariate balance test for the dose random assignment. The difference reported in column 3 is a between-school difference: the difference between (i) treatment schools assigned to high math, low English, and (ii) treatment schools assigned to low math, high English. This test shows no statistically significant differences, as we would expect after successful random assignment.

[3] Our funder, the Education Endowment Foundation (EEF), requires that all experiments have an independent evaluator. Thus, the independent evaluator, the National Foundation for Educational Research (NFER), also reported on the experiment (Worth et al. 2017). This paper and the NFER report were undertaken separately with different authors and analysis plans, and the two were planned as such from before the experiment began. Additionally, under the EEF's rules, the random assignment procedures were carried out by NFER. The procedures were designed jointly by the authors of this paper and NFER.

Table 1
Pretreatment Characteristics

		Difference in Means	
	Sample Mean (SD) (1)	School Assigned to Treatment − Control [*p*-Value] (2)	School Assigned to Math Dept. High Dose − Math Low [*p*-Value] (3)
Prior math score	.007	−.006	−.037
	(.998)	[.872]	[.489]
Prior English score	.006	−.029	−.023
	(.999)	[.477]	[.695]
Female	.487	−.020	.016
		[.279]	[.369]
IDACI	.276	.031	.023
	(.171)	[.068]	[.368]
Free school meals	.398	.019	.018
		[.382]	[.613]
Birth month (1–12)	6.569	−.034	.060
	(3.419)	[.290]	[.207]
London school	.162	.028	.102
		[.703]	[.441]
Differences jointly zero		[.323]	[.246]

NOTE.—For each pretreatment characteristic, col. 1 reports the study sample mean (standard deviation), with 28,704 student observations, the full sample. Column 2 reports the treatment minus control difference in means and the *p*-value for the null that the difference is zero. Differences and *p*-values come from a regression of the pretreatment characteristic on an indicator for treatment and randomization block fixed effects. The standard errors allow for clustering at the school level. The bottom row reports on a joint test that all of the treatment-control differences are simultaneously zero. Column 3 reports the difference in means between treatment schools assigned to (i) high-dose math and low-dose English and those assigned to (ii) low-dose math and high-dose English. Differences and *p*-values are estimated as in col. 2, except that the sample is limited to treatment schools. IDACI = Income Deprivation Affecting Children Index.

2. Teachers

For each treatment school, we randomly assigned teachers to different roles in the program. One-third were assigned to be "observers" who visited other teachers' classrooms and scored the peer they watched. Observers were not paired with observees and could observe either math or English lessons. One-third were assigned to be "observees" whose teaching would be observed and scored by the observers. And the final one-third were assigned to take both observer and observee roles. Role random assignment was within department—that is, within school-by-subject blocks—and independent of the other randomizations.

In table 2 we test for balance across randomly assigned roles. We do not have data on the teachers themselves, so in table 2 we compare characteristics of students in the classrooms of observer and observee teachers. To emphasize the difference, despite using the same student covariates, in table 1 we test for differences between schools assigned to different conditions, while in table 2 we test for differences between teachers within departments.

Table 2
Pretreatment Characteristics by Teacher Role

	Difference in Means			
	Experiment Year 1		Experiment Year 2	
	Observee − Observer [*p*-Value] (1)	Both Roles − Observer [*p*-Value] (2)	Observee − Observer [*p*-Value] (3)	Both Roles − Observer [*p*-Value] (4)
Prior math score	−.052	−.009	−.139	.059
	[.297]	[.850]	[.167]	[.555]
Prior English score	−.053	.004	−.128	.030
	[.284]	[.927]	[.191]	[.765]
Female	−.027	.002	−.008	−.002
	[.017]	[.849]	[.675]	[.937]
IDACI	.006	.005	.019	.006
	[.083]	[.143]	[.005]	[.366]
Free school meals	.027	.014	.041	−.004
	[.009]	[.189]	[.048]	[.821]
Birth month (1–12)	.057	.071	.115	.089
	[.346]	[.285]	[.224]	[.332]
Differences jointly zero	[.082]		[.246]	

NOTE.—For each pretreatment characteristic, col. 1 reports the difference in means: (i) the mean student characteristic of students assigned to observer teachers minus (ii) the mean for observee teachers, with *p*-values in brackets for the null that the difference is zero. Column 2 similarly reports the difference in means: (i) observer teachers minus (iii) dual-role teachers. The estimates in cols. 1 and 2 come from regressing each pretreatment characteristic on an indicator for observee and an indicator for both-role and randomization block fixed effects. Standard errors allow for clustering at the teacher level. The bottom row reports on a joint test that all of the differences in cols. 1 and 2 are simultaneously zero. The estimation sample for cols. 1 and 2 is limited to year 1 of the experiment. For cols. 3 and 4 we repeat the same estimation process as used in cols. 1 and 2, except that in cols. 3 and 4 the estimation sample is limited to year 2. IDACI = Income Deprivation Affecting Children Index.

Teachers were randomly assigned to roles in early October 2014—at the beginning of the experiment's first school year but after students had been assigned to classes and teachers. In columns 1 and 2 of table 2 we use only data from that first year, 2014–15. Unlike table 1, here we do see more differences than we would expect. Compared with students of observer teachers, students of observee teachers are less likely to be female and have more exposure to poverty in their homes and neighborhoods. However, students are relatively similar on prior achievement. Also, covariates are well balanced comparing observer and dual-role teachers. The joint test of all 12 differences has a *p*-value of .082. Below, after presenting the observer-observee results, we discuss interpretation of those results given the imbalance in table 2 and provide some relevant robustness tests.

The second year of the experiment, 2015–16, differed from the first year in some respects. Those differences created opportunities for endogenous sorting of teachers and students, and thus potential bias. Schools had some new hires who needed role assignments; 2015–16 new hires are 12.2% of

our sample. Some new hires were randomly assigned to roles, but others took the role of the departing teacher they replaced.[4] All returning teachers kept the same randomly assigned role for both years 1 and 2. However, unlike year 1, teachers' roles were known when students were assigned to returning teachers' classes. In columns 3 and 4 of table 2 we use only data from that second year, 2015–16. The differences in exposure to poverty are slightly stronger, although there is no longer a gender difference. While the differences in prior achievement are not statistically significant, they are more than twice as large as year 1.

Finally, throughout the paper results comparing observers, observees, and dual-role teachers—including tables 2 and 6—are based on 33 treatment schools. Teachers were randomly assigned in all 41 treatment schools, but eight schools later withdrew their consent for the use of their class rosters. Those rosters are the only way to link teachers to students.[5] Still, all 41 treatment schools (and all 41 controls schools) are used for estimating overall school-level effects and dosage effects.

B. Description of the Treatment

The treatment, in short, is peer classroom observations among coworkers teaching in the same school. As described above, teachers were randomly assigned to either observe or to be observed. Each classroom observation was scored using a detailed rubric based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (2007; hereafter, FFT) and lasted approximately 15–20 minutes. The stated goal was 6 or 12 observations per observee teacher per year, where 6 or 12 was randomly assigned as described above. Teachers were encouraged to meet after observations to discuss feedback and share strategies for improvement.

The FFT rubric is widely used by schools and in research (e.g., Kane et al. 2011, 2013; Bacher-Hicks et al. 2019). The rubric is divided into four "domains"—classroom environment, instruction, planning, and assessment—with several "standards" within each domain. In the current experiment, classroom observations used only the classroom environment and teaching domains, which are measured while watching teachers teach. In table 3 the left-hand column lists the 10 standards on which teachers were evaluated. For each standard, the rubric includes descriptions of what observed behaviors should be scored as "highly effective," "effective," "basic," and "ineffective" teaching. In table 3 we reproduce the descriptions for effective teaching as an example. The full rubric is provided in appendix B.

[4] We do not have data on which new hires assumed a role and which were randomly assigned.

[5] As explained below, teacher-student class rosters were provided to the research team by each school directly. Student test scores come from the National Pupil Database, which does not link teachers to students.

Table 3
Rubric Standards and Associated Description of "Effective"

	Description
Domain 1, classroom environment:	
1a. Creating an environment of respect and rapport	Classroom interactions, both between teacher and students and among students, are polite and respectful, reflecting general warmth and caring, and are appropriate to the cultural and developmental differences among groups of students.
1b. Establishing a culture for learning	The classroom culture is characterized by high expectations for most students and genuine commitment to the subject by both teacher and students, with the teacher demonstrating enthusiasm for the content and students demonstrating pride in their work.
1c. Managing classroom procedures	Little teaching time is lost because of classroom routines and procedures for transitions, handling of supplies, and performance of nonteaching duties, which occur smoothly. Group work is well organized, and most students are productively engaged while working unsupervised.
1d. Managing student behavior	Standards of conduct appear to be clear to students, and the teacher monitors student behavior against those standards. The teacher response to student misbehavior is consistent, proportionate, appropriate, and respects the students' dignity.
1e. Organizing physical space	The classroom is safe, and learning is accessible to all students; the teacher ensures that the physical arrangement is appropriate for the learning activities. The teacher makes effective use of physical resources, including computer technology.
Domain 2, teaching:	
2a. Communicating with students	Expectations for learning, directions and procedures, and explanations of content are clear to students. Communications are accurate as well as appropriate for students' cultures and levels of development. The teacher's explanation of content is scaffolded, clear, and accurate and connects with students' knowledge and experience. During the explanation of content, the teacher focuses, as appropriate, on strategies students can use when working independently and invites student intellectual engagement.
2b. Using questioning and discussion techniques	Most of the teacher's questions elicit a thoughtful response, and the teacher allows sufficient time for students to answer. All students participate in the discussion, with the teacher stepping aside when appropriate.

Table 3 (*Continued*)

	Description
2c. Engaging students in learning	Activities and assignments, materials, and groupings of students are fully appropriate for the learning outcomes and students' cultures and levels of understanding. All students are engaged in work of a high level of rigour. The lesson's structure is coherent, with appropriate pace.
2d. Use of assessment	Assessment is regularly used in teaching, through self-assessment or peer assessment by students, monitoring of progress of learning by the teacher and/or students, and high-quality feedback to students. Students are fully aware of the assessment criteria used to evaluate their work and frequently do so.
2e. Demonstrating flexibility and responsiveness	The teacher promotes the successful learning of all students, making adjustments as needed to lesson plans and accommodating student questions, needs, and interests.

NOTE.—Adapted from *Enhancing Professional Practice: A Framework for Teaching* (Danielson 2007) for the current experiment.

The FFT and other similar observation rubrics and protocols were not explicitly designed as a tool to improve student achievement scores, as we measure in this experiment. Nevertheless, existing evidence is consistent with expecting positive effects of this "treatment." First, several studies now find a similar, if moderate, positive correlation between observation scores and student test scores, including some settings where students are randomly assigned to teachers (Kane et al. 2011, 2013; Garrett and Steinberg 2015; Araujo et al. 2016; Bacher-Hicks et al. 2019). Second, as cited in the introduction, a growing number of (quasi)-experimental studies document positive effects on student achievement of programs where teachers are observed and scored using FFT or similar rubrics (Taylor and Tyler 2012; Steinberg and Sartain 2015; Garet et al. 2017; Briole and Maurin 2019).

In addition to the FFT-based assessments of teaching quality, we also asked observers to record other relatively objective data on teaching practices. For example, how often—never, some of the time, or all of the time—the teacher lectured the whole class, had students work in small groups, or taught students one on one.

Teachers were provided training on the FFT rubric and other aspects of the program, primarily in person but supplemented with phone and email conversations. Treatment schools were also given a few tablet computers to facilitate observations. Observers could access the complete FFT rubric, record scores, and make notes during their observations.[6] The centrally

[6] The tablets were Apple iPads, and the software was created by RANDA Solutions.

stored database of observations allowed the research team to monitor progress of individual schools and contact those who were clearly lagging. However, the specific schedule and pace of observations was left to each school to determine.

C. Evaluation and Classroom Observation in English Secondary Schools

Classroom observations of teaching are certainly not new to the schools we study, and observations were likely occurring in control schools during the experiment. However, the treatment observations had three (potentially) first-order features distinct from business as usual in English secondary schools: observation by peer teachers, observations based on an FFT-style rubric, and simply more observations regardless of who or what rubric.

In our conversations with study schools, most reported that some form of class observations were part of their normal business. In contrast to the treatment peer observations, these school-initiated observations were conducted by school leaders, unstructured, and much rarer. The average teacher would be observed perhaps once per year and often less than annually. Moreover, the frequency of observations was curtailed partly by union opposition, sometimes codified into rules limiting observations. Consistent with this description of limited status-quo observation, treatment teachers reported appreciating that the program included more frequent observations and observations from peers instead of school leaders.

Beyond school-initiated efforts, classroom observations occur for two other reasons. First, observations are part of the formal induction and assessment process for novice teachers, known as "newly qualified teacher" (NQTs) or the "NQT year." An NQT might be observed as many as six times, but the teachers are typically only NQTs for one year.

Second, observations are part of England's formal performance evaluation process for schools. The school inspection process, conducted by the Office of Standards in Education, Children's Services and Skills (Ofsted), does include classroom observations among many other forms of evaluation. However, Ofsted's classroom observations are not salient to individual teachers. During a school inspection, Ofsted observers visit several teachers' classes, but far from all teachers' classes; the goal of Ofsted's observations is to make overall assessments of teaching in the school, not of each individual teacher. Moreover, a typical school might be visited by Ofsted only every 3 or 4 years. In short, the chances of a given teacher in a given year being observed by Ofsted are low, and there would be little individual consequence of the results.[7]

[7] Ofsted and the Department for Education also set expectations and guidelines for each school's own self-evaluation practices. Classroom observations, broadly speaking, should be part of each school's plans, but Ofsted and the Department

D. Sample and Data

Our study sample is composed of 82 schools, more than 28,000 students with GCSE math and English test scores, and approximately 1,300 teachers. We initially contacted almost all high-poverty public (state) secondary schools and invited them to participate in the study.[8] School performance levels (test scores) were not used as a criterion for inviting schools. Of the invited schools, 92 volunteered to participate (8.5%). We randomized 82 schools to treatment and control, after 10 volunteers dropped out before randomization.[9] The schools initially invited were intentionally selected to have high poverty rates. These characteristics are reflected in the study sample, as shown in table 1, column 1. Nearly 40% of students are (or ever have been) eligible for free school meals, substantially higher than the national average.

Nearly all of the data for our analysis come from the UK government's National Pupil Database (NPD). These administrative data include student-level records with math and English GCSE scores, prior test scores (Key Stage 2), demographic and other characteristics of students, and their school. The NPD data are sufficient for our intent-to-treat (ITT) estimates of average treatment effect.

We add to the NPD data in two ways. First, the data recorded during peer observations allow us to measure participation, for example, the number of observations completed. The data also include observation scores, but we do not use those scores in this paper. Second, each school provided class rosters, which link teachers to students but are not available in the NPD. The rosters list each school's math and English teachers and the specific year 10 and 11 students assigned to those teachers. We link rosters to NPD data using unique student IDs. Schools provided rosters before random assignment. We use these linked teacher-student roster data in our analysis of teacher roles and to examine QTEs by individual teacher performance.

Our estimates are not threatened by attrition, at least not in the first-order sense of attrition. The NPD data include the universe of students and schools. Thus, even if a school chose to withdraw from the study, we observe outcomes and can still include the school in our analysis. If, however, treatment induced students to move to different schools at a rate higher (lower) than control schools, those movements would be relevant to the interpretation of our results. Treatment effects on student school switching seems unlikely.

---

for Education do not require a specific minimum number, type of observer, or criteria for what should be evaluated in the observation. Moreover, until recently there was a rule limiting observations to no more than three per year.

[8] We excluded, ex ante, boarding schools, single-gender schools, and schools in select geographic areas where the funder was conducting different interventions. The final list invited was 1,097 schools.

[9] One additional school in Wales volunteered, making 93 total, but was excluded because the NPD only covers England.

Treatment effects on teachers switching schools may be more plausible. Our analysis of differences between teacher roles uses teacher's class rosters provided at the beginning of each school year, in the spirit of ITT, and we show that those results are similar in each year separately.

III. Methods

Our analysis of the experimental data follows conventional experimental methods. To begin, we estimate the difference between average student GCSE scores in treatment and control schools by fitting the following regression specification:

$$Y_{imt} = \delta T_s + \pi_b + X_i\beta + \varepsilon_{imt}, \quad (1)$$

where Y_{imt} is the GCSE score for student i in subject m (math or English) taken in year t (2015 or 2016). Student scores, Y_{imt}, are standardized (mean = 0, standard deviation = 1) by subject and year within our analysis sample. The indicator $T_s = 1$ for all schools s randomly assigned to the treatment, and π_b represents fixed effects for the eight blocks b within which random assignment occurred.[10] The vector X_i includes the student characteristics measured pretreatment and listed in table 1, most notably prior achievement scores in math and English, plus indicator variables for year t and for subject m. We report cluster-robust standard errors where the clusters are schools s, the unit at which treatment is assigned.[11]

Fitting specification (1) returns ITT effects. We also report treatment-on-the-treated (TOT) estimates where T_s is replaced with an indicator equaling 1 if the school actually implemented the peer observation program, and we instrument for that endogenous treatment indicator with the randomly assigned T_s. Because the latent characteristic "implemented" is not binary, we show a few different alternatives for the endogenous treatment indicator.

A causal interpretation of our estimate $\hat{\delta}$ requires the conventional experimental identification assumption: in the absence of the experiment, students in treatment and control schools would have had equal GCSE scores at expectation. The balance in pretreatment covariates (table 1) across treatment and control schools is consistent with this assumption. However, the imbalance in the teacher role experiment (table 2) does not threaten this assumption. Table 1 tests for between-school differences, while table 2 tests for between-teacher within-department differences. When we turn to interpreting mechanisms in section V, imbalance across roles is potentially relevant.

[10] Students are nested in schools, $s = s(i)$, and schools are nested in randomization blocks, $b = b(s)$. To streamline the presentation we use the simple s and b subscripts.

[11] Our main estimates pool subjects. Students' math and English score errors are also likely correlated. Since students are nested within schools, clustering by school is identical to clustering by school and student.

To examine how the number of peer observations contributes to treatment effects, we use the high- and low-dose experimental conditions. We add an indicator $H_m = 1$ if department m was randomly assigned to the high-dose condition, $H_m = 0$ if department m was low dose, and $H_m = 0$ for all control schools:

$$Y_{imt} = \delta T_s + \gamma H_m + \pi_b + X_i\beta + \varepsilon_{imt}. \quad (2)$$

Thus, the coefficient γ measures the added (reduced) treatment effect above (below) δ. The identifying assumption for $\hat{\gamma}$ is similar: in the absence of the experiment, students in high- and low-dose departments would have had equal GCSE scores at expectation. Table 1, column 3, supports this assumption.

We also estimate differences in student outcomes by their teacher's role in the peer observation: observer, observee, or both. Using observations from treatment schools, we fit the following specification:

$$Y_{imt} = \alpha_1 \text{VEE}_{j(imt)} + \alpha_2 \text{BOTH}_{j(imt)} + \omega_{sm} + X_i\beta + \nu_{imt}, \quad (3)$$

where $\text{VEE}_{j(imt)}$ is an indicator equaling 1 if student i's teacher j in year t for subject m was randomly assigned to be an observee, and similarly $\text{BOTH}_{j(imt)}$ equals 1 if the teacher was assigned to both roles. The omitted category is when the teacher was assigned to be an observer. The ω_{sm} represents fixed effects for department; random assignment of roles occurred within departments, that is, school-by-subject cells. Recall that the sample for estimating specification (3) is a subset of the treatment sample for specifications (1) and (2) because we lack teacher-student class rosters for eight treatment schools. We report cluster-robust standard errors where the clusters are teachers j, the unit at which treatment is assigned.

Causal interpretation of the role differences, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, requires a between-teacher identifying assumption: in the absence of the experiment, students in observee, observee, and "both-role" teachers' classrooms would have had equal GCSE scores at expectation. The relevant pretreatment covariate tests are shown in table 2. We return to the interpretation of the role differences after presenting the results.

Last, we test whether the treatment effect is larger (smaller) for lower-performing teachers compared with higher-performing teachers. We do not have data on individual teacher performance in any preexperiment years.[12] Instead, we adopt a QTE approach. First, we estimate a value-added score for each teacher, $\hat{\mu}_j$, which measures a teacher's contribution to her assigned students' GCSE scores during the experiment. We follow the methods described in Kane and Staiger (2008) and Chetty, Friedman, and Rockoff

[12] The barrier is that we do not have class roster data linking teachers to students for the preexperiment years, only for the experiment years. Thus, we cannot calculate value-added scores prior to the experiment.

(2014a).[13] Second, we estimate the treatment-control difference in $\hat{\mu}_j$ at a given percentile of the teacher value-added distribution, where, critically, percentiles are determined separately for the treatment distribution and for the control distribution.

We use the unconditional quantile regression method proposed by Firpo, Fortin, and Lemieux (2009). The regression specification is as follows:

$$\left[\tau - 1\left\{\mu_j \leq q_\tau\right\}\right]\left[f_{\mu_j}(q_\tau)\right]^{-1} = \delta^\tau T_s + \pi_b^\tau + \epsilon_j^\tau, \tag{4}$$

where the dependent variable on the left is the influence function (IF) for the τth quantile, $\text{IF}(\mu_j; q_\tau, F_{\mu_j})$, T_s is the treatment indicator, and π_b^τ are the randomization block fixed effects. Firpo, Fortin, and Lemieux (2009) detail the properties of this IF-based estimator, which are straightforward in this randomly assigned binary treatment case. For inference we construct cluster bootstrap–based 95% confidence intervals, where resampling is at the school level. After presenting the QTE results below, we discuss what causal claims can be made on the basis of the results under different assumptions.

IV. Results

A. Effect of the Peer Evaluation Program

Teacher peer evaluation, with the program features described in section II, produced educationally and economically meaningful improvements in student learning, most likely by boosting teachers' contributions to their students' learning. During the two treatment years, students in treatment schools scored 0.06–0.07σ higher (ITT), on average, than their counterparts in control schools. In treatment schools that took up the program, the benefit was at least 0.09σ (TOT).

Estimates of the differences between treatment and control schools are shown in table 4. The simplest treatment effect estimate, 0.056σ, is shown in column 1: ITT, pooling subjects, and controlling only for randomization block fixed effects. That simple estimate is not precisely estimated, however. In column 2 we add pretreatment controls to improve precision. Under the assumption that random assignment balanced expected potential outcomes, both column 1 and column 2 are consistent estimates for the causal effect of treatment. While the point estimates are somewhat different, 0.056σ and 0.073σ, we cannot reject the statistical null hypothesis that they are the same.

Improvements of 0.06σ–0.07σ in math and English are educationally and economically significant. The most likely mechanisms for these results, which

[13] Our data include only two school years, so we adopt the assumption that value added is fixed over those two years. The outcome measure Y_{imt} is student GCSE scores taken at the end of year 11. We do not have a prior year test score, $Y_{im,t-1}$; thus, instead of the more common lagged dependent variable approach, we use student fixed effects in the residualization step. In sec. V we discuss how the student fixed effects approach affects interpretation.

Table 4
Effect of Teacher Peer Observation Program on Student Achievement Scores

	A. Intent-to-Treat Estimates		
	(1)	(2)	
School randomly assigned to treatment	.056	.073*	
	(.040)	(.032)	
Pretreatment covariates		✓	
Adjusted R^2	.022	.343	
Observations	56,148	56,148	
	B. Treatment-on-the-Treated Estimates under Alternative Definitions of Treatment Take-Up		
	(3)	(4)	(5)
School completed at least one peer observation	.093*		
	(.039)		
School completed at least 10% of suggested observations		.133*	
		(.057)	
At least 50% of teachers participated once or more often			.156*
			(.071)
Pretreatment covariates	✓	✓	✓
Adjusted R^2	.345	.344	.342
Observations	56,148	56,148	56,148
First-stage *F*-statistic excluded instrument	160.8	65.4	38.3

NOTE.—Each column reports results from a separate least squares (panel A) or two-stage least squares (panel B) regression, with student-by-subject observations. The dependent variable is student math or English General Certificate of Secondary Education score in student standard deviation units. All specifications include randomization block fixed effects and an indicator for math observation. Pretreatment covariates include the characteristics listed in table 1 and an indicator for cohort 1. When a pretreatment covariate is missing, we replace it with zero and include an indicator variable equaling 1 for missing on the given characteristic. For the instrumental variables estimates in panel B, the row headers describe the endogenous treatment indicator variable, which is instrumented for with the randomly assigned treatment condition indicator. Cluster-robust standard errors are reported in parentheses, with clustering at the school level.
* $p < .05$.

we discuss in more detail below, are mechanisms that operate through teachers' causal contributions to student test scores. Improving a teacher's contribution by 0.06σ–0.07σ would move her up perhaps one-fifth to one-quarter of a standard deviation in the teacher performance distribution, as measured by teacher contributions to student test scores.[14] Such improvements in teacher

[14] Slater, Davies, and Burgess (2011) estimate the standard deviation of teacher contributions to GCSE (value-added) scores is 0.272σ. This estimate comes from English secondary schools and GCSE courses, as in our current study, although the sample in Slater, Davies, and Burgess (2011) is broader. Judged against this 0.272 estimate, the treatment effects would be one-fifth to one-third of a teacher standard deviation. The 0.272 estimate may be larger than other estimates (e.g., from US elementary and middle schools) in part because students spend 2 years

performance are large but not unprecedented in the literature. Taylor and Tyler (2012) find improvements of 0.05σ–0.11σ from a relatively similar peer evaluation program. Jackson and Makarin (2018) report improvements of 0.06σ–0.09σ in an experiment where teachers were provided with high-quality math lesson plans. Finally, our estimate of 0.06σ–0.07σ is roughly similar to the performance improvements made by new teachers in the first 3–5 years on the job (for a recent review, see Jackson, Rockoff, and Staiger 2014).

As is common in field experiments, some schools randomly assigned to the treatment and encouraged to implement the new peer evaluation program nevertheless chose not to participate or participated relatively little. In columns 3–5 of table 4 we report rough bounds for the treatment effect on the treated. Thirty-four of the 41 treatment schools (83%) completed at least one peer observation. Thus, a simple lower bound on the TOT estimate is $\hat{\delta}$ divided by 0.83. Column 3 of table 4 formalizes this estimate using two-stage least squares, where the endogenous treatment indicator equals 1 if the school completed at least one peer observation. (First-stage estimates for columns 3–5 are provided in table A2; tables A1–A6 are available online.) Any upper-bound TOT estimate is rougher because we must choose some cutoff for a more restrictive definition of "implemented" the treatment. In column 4 of table 4 the endogenous treatment indicator equals 1 if the school completed 10% or more of the peer observations they were originally asked to conduct.[15] Of the 41 schools, 25 met the (admittedly somewhat arbitrary) condition of 10% or more. If we scale up the ITT by this stricter first stage, the implied TOT would be 0.13σ. In column 5 the endogenous indicator equals 1 if more than half of teachers participated in at least one observation, as observer or observee. If we scale up the ITT by the implied first stage, we get an estimate of 0.16σ. In short, we believe that a plausible range for the TOT effect is roughly 0.09σ–0.16σ. In the next subsection we return to the question of whether treatment effects depend on the number of peer observations.

The results in table 4, and throughout most of the paper, pool subjects and years. Pooling simplifies the presentation of results and improves precision. In table A3 we report estimates separately by subject and student cohort. The point estimates are slightly larger for English GCSE scores and in the first year of the treatment, but differences across subject and year are not statistically significant at conventional levels.

---

with their GCSE teacher. For a general summary of estimates on the teacher value-added distribution, see Jackson, Rockoff, and Staiger (2014) and Hanushek and Rivkin (2010). However, many estimates of the teacher value-added distribution come from elementary and middle schools, and the variation may be greater or smaller in (English) secondary schools.

[15] We asked schools to complete 6 or 12 observations per year for each observee teacher. Whether 6 or 12 was determined by the department-level dosage assignment is discussed and analyzed in the next subsection.

B. Treatment Effects and the Number of Peer Observations

A first-order design feature of a teacher observation program, like the one we study, is the number of classroom observations conducted. Our treatment effect estimates, discussed in the previous subsection, are effects for the average treatment school. The corresponding number of observations completed by the average treatment school is 2.27 per observee teacher per year (standard deviation = 2.67). The natural follow-up question is, Would the estimated average treatment effect be larger or smaller than 0.073σ if the number of observations conducted were larger or smaller?

Our answer comes from direct experimental variation in the number of observations. Recall that within each treatment school, one of the two departments (math or English) was randomly assigned to a low-dose condition of 6 observations per observee per year, and thus the other of the two departments (English or math) was randomly assigned to a high-dose condition of 12 observations. The average number of observations actually conducted was 1.6 and 2.9 in the normal- and double-dose conditions, respectively (see the "first-stage" results in table A4). Actual observations were certainly short of what teachers were initially asked to do, but the difference in actual observations caused by the dosage random assignment was still nearly a doubling of observations.

That experimental doubling of observations did not increase the treatment effect. In table 5, column 1, we show estimates from a specification identical to table 4, column 2, except that we have added a right-hand-side indicator

Table 5
Number of Peer Observations and Treatment Effects

	(1)	(2)
Treatment school	.074*	
	(.033)	
High-dose department	−.002	−.002
	(.020)	(.025)
Pretreatment covariates	✓	
Randomization block fixed effects	✓	
Student fixed effects		✓
Adjusted R^2	.343	.338
Observations	56,148	29,456

NOTE.—Each column reports results from a separate least squares regression, with student-by-subject observations. The dependent variable is student math or English General Certificate of Secondary Education score in student standard deviation units. For col. 1, the specification includes the pretreatment covariates listed in table 1, plus an indicator for math observation and an indicator for cohort 1. When a pretreatment covariate is missing, we replace it with zero and include an indicator variable equaling 1 for missing on the given characteristic. For col. 2, these additional covariates are replaced by a student fixed effect. Column 2 uses only observations from treatment schools. Cluster-robust standard errors are reported in parentheses, with clustering at the school level.

* $p < .05$.

equaling 1 if the department was randomly assigned to the high-dose condition (the new indicator is 0 for all control cases). The estimated effect is close to zero, -0.002σ, and far from statistically significant. This test, though binary, suggests little covariance between treatment effects and number of observations over the range of 1.6–2.9 observations. As a robustness check, in column 2 we estimate a student fixed effects version of equation (2).

The lack of dose effect is not inconsistent with prior evidence. Taylor and Tyler (2012), Steinberg and Sartain (2015), and Briole and Maurin (2019) report quasi-experimental estimates for similar classroom observation interventions where the number of observations per year was four, two, and one, respectively. Over this one to four range, which includes our 1.6–2.9 contrast, the estimated effects varied little: 0.052σ, 0.054σ, and 0.045σ, respectively, in math during the year when observations occurred. While our results are consistent with a prior based on these estimates, our contribution is identification by random assignment.

C. Treatment Effects and Teacher Roles

Another first-order feature of a teacher observation program is deciding who should be observed (the observees) and by whom (the observers). By randomly assigning these teacher roles, our experiment contributes distinctively new evidence to the literature on teacher peer evaluation. First, the random assignment of roles is itself a feature of the treatment intervention and is thus relevant to a discussion of the potential mechanisms that gave rise to the overall effect of 0.073σ. We return to this topic in section V. Second, more directly, the role experiment allows us to estimate treatment effects separately for observer and observee teachers.

Results of the role experiment are reported in table 6. The students of observer teachers scored 0.057σ higher than the students of observee teachers (col. 2, which controls for pretreatment observables). A difference of 0.057σ is large in this context: about one-fifth of the standard deviation in teacher performance, as measured with student test scores, and nearly four-fifths of the overall treatment effect of 0.073σ. However, while 0.057σ is substantively large, it is not statistically significantly different from zero. The 95% confidence interval runs from a difference of 0.016σ favoring observees to 0.130σ favoring observers. We have less power here than when testing for the overall treatment effect. Moreover, the covariate balance tests, discussed in section II, suggest caution when making inferences about the causes of these observer-observee differences.

In short, we cannot draw strong conclusions about the observer-observee differences. We cannot rule out the conclusion that the treatment effects for observers were equal to the effects for observees, nor can we rule out that effects were larger for observers. The remainder of this section further examines the magnitude, precision, and robustness of the observer-observee results. In the end, the limits on strong conclusions remain. Still, if we keep

Table 6
Treatment Effects by Teacher Role

	Experiment Year					
	Pooled			Year 1	Year 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher role (relative to observer):						
Observee	−.136*	−.057	−.075*	−.083+	−.017	−.050
	(.063)	(.037)	(.037)	(.049)	(.047)	(.051)
Both roles	−.010	−.015	−.025	−.050	.024	.012
	(.060)	(.035)	(.036)	(.049)	(.040)	(.043)
Pretreatment covariates		✓	✓	✓	✓	✓
Year 1 teacher controls			✓			✓
Adjusted R^2	.069	.371	.372	.323	.478	.479
Observations	15,077	15,077	15,077	8,687	6,390	6,390

NOTE.—Each column reports results from a separate least squares regression, with student-by-subject observations. "Pooled" combines experiment year 1 and experiment year 2 samples. The dependent variable is student math or English General Certificate of Secondary Education score in student standard deviation units. All specifications include department fixed effects (i.e., the school-by-subject randomization blocks) and an indicator for math observation. Pretreatment covariates include the characteristics listed in table 1 and an indicator for cohort 1. When a pretreatment covariate is missing, we replace it with zero and include an indicator variable equaling 1 for missing on the given characteristic. Year 1 teacher controls include indicators for role: observer, observee, both roles, and no role. Cluster-robust standard errors are reported in parentheses, with clustering at the teacher level.

+ $p < .10$.

* $p < .05$.

the limitations in mind, the observer-observee results are a useful input to the discussion of costs and potential mechanisms in sections V and VI.

A primary goal of randomly assigning teacher roles was to measure any differences in treatment effects caused by a teacher role. That causal interpretation of the 0.057σ observer-observee difference relies on the assumption that random assignment successfully balanced potential outcomes. The conventional test of that assumption in table 2 shows some unexpected imbalance in pretreatment observables. That same imbalance can also be seen comparing columns 1 and 2 of table 6, where column 2 controls for pretreatment covariates and column 1 does not. Controlling for differences in observables is straightforward, but as always the concern is that observable differences suggest scope for unobserved differences. In the end, this imbalance should add caution to a strong causal interpretation of the differences between roles. In the remainder of this subsection we discuss some additional results relevant to assessing the scope of potential bias. However, in the end these additional results cannot rule out all potential bias, and so our caution remains. The relevant source of bias may be shared by all of the checks, even as those checks differ in other ways.

The first result comes from simply limiting the estimation sample to the first year of the experiment, 2014–15. Recall from section II that the main random assignment of roles occurred at the start of the 2014–15 school year after teachers and students had already been assigned to classes but that year 2 created new opportunities for endogenous sorting of teachers (roles) to

students. And in table 2 we saw that observable pretreatment differences were smaller in year 1. All of which suggests year 1 might be a preferable, but less precise, test of role effects. In year 1, the observer-observee difference is 0.083σ (table 6, col. 4) and marginally different from zero ($p = .088$). While the year 1 point estimate of 0.083σ is larger than the pooled 0.057σ, we cannot reject the null that they are equal.

An additional feature of year 2, which we have not discussed so far, further complicates the analysis of teacher roles but adds clarity to the results. Students who took the math GCSE exam in 2014–15—the first year of the experiment—had been taught by one treated math teacher. But students who took the math GCSE exam in 2015–16—the second year of the experiment—had been taught by (possibly) two different treated math teachers, their 2014–15 year 10 teacher and their 2015–16 year 11 teacher. (The same holds for English teachers and exams.) Thus, the potential endogenous student-teacher sorting in year 2 includes sorting based on the prior teacher's role. In the end, controlling for prior teacher's role makes a difference in our estimates. First, column 5 of table 6 uses only observations from year 2 of the experiment, but the specification is otherwise identical to columns 2 and 4. Then in columns 3 and 6 we add controls for prior teacher's role.[16] After adding these controls, the observer-observee differences are much more consistent.

In the appendix we show results from two further robustness tests. In the first we use only within-student variation: differences, for a given student, in her math teacher's role and her English teacher's role. The student fixed effects results are consistent with our main estimates: point estimates favoring observers, but generally no statistically significant differences between observers and observees. Even if there was between-student sorting across teacher roles, with the student fixed effects, threats to identification would require sorting based on differences between math and English potential outcomes for a given student.

The second robustness test uses the fact that role assignment occurred within departments in treatment schools. We can thus estimate the degree of pretreatment covariate (im)balance for each school and then see whether treatment effects covary with the estimate of (im)balance. In schools with relatively low imbalance, the pattern of effects mirrors the main estimates: small differences that are not statistically significant. However, we cannot say what made some schools relatively low imbalance and others high imbalance, and so we are cautious about generalizing the results.

[16] These controls are indicator variables analogous to $\text{VEE}_{j(imt)}$ and $\text{BOTH}_{j(imt)}$ in eq. (3) but defined on the basis of the teacher student i was assigned in $t - 1$ for subject m. We also include an indicator equaling 1 if that $t - 1$ teacher was none of observer, observee, or dual role. See the discussion of these "nonparticipating teachers" below. An alternative to controlling for prior teacher's role is to control for prior teacher fixed effects; this yields similar results, although the observer-observee point estimates are larger (results are provided in table A5).

To this point we have been discussing differences in treatment effects between observers and observees. What can we say about treatment effect levels across teacher roles? The average treatment effect level across all teachers is 0.073σ. If 0.073σ was simply a weighted average of effects for observers and observees, then the effect level for observers would be about 0.097σ, and that for observees would be 0.040σ. However, some of the teachers in treatment schools did not participate in the role experiment. The 0.073σ estimate is a weighted average of observers, observees, and these other nonparticipating teachers. The effects for observers and observees will be larger than 0.097σ and 0.040σ as long as the treatment effect for nonparticipants is smaller than the treatment effect for participants. We estimate that the effect for observers was 0.120σ and that for observees was 0.063σ; these estimates, and the accompanying assumptions, are detailed in appendix C.

Finally, we note that one consistent result—across estimates in table 6 and elsewhere—is that the point estimates always show observers with better outcomes than observees. This suggests some confidence in ruling out the conclusion that effects were larger for observees. However, to reiterate, we do not rule out that the treatment effects were the same for observers and observees; results for testing the null hypothesis of no difference are much less consistent than the direction of the point estimates.

V. Mechanisms

What mechanisms might have created the gains in treatment schools? One conventional hypothesis is that teachers worked harder while their performance was being scored. There were no explicit incentives or consequences to motivate extra effort, but teachers may have been motivated by social pressure or career concerns. However, this mechanism alone would not explain the 0.073σ gain. First, extra effort for 20 minutes, two or three times a year, would not change student test scores. Second, only observee teachers were scored, but treatment effects were just as large, if not larger, for the observer teachers.

A second potential mechanism is that teachers' skills improved. Learning new skills requires effort, but it also requires knowing where to direct that effort. First, being scored with a rubric creates new information likely helpful in deciding where to direct effort. Peer observations create personalized feedback—which is rare for teachers—including a teacher's own relative strengths and weaknesses across the rubric's 10 different skill areas (table 3). Teachers may also learn or infer information about how their own performance compares with that of other teachers.[17] Additionally, the rubric itself (implicitly) provides practical advice on what an "ineffective" teacher should do differently

[17] The experiment did not require, but also did not prohibit, that teachers share their scores with each other. Also, the rubric itself is a normative statement, suggesting what teachers should and can do.

to become "effective" in the future; that advice comes from the rubric's concrete descriptions of teacher behavior and actions associated with each score. Second, peer evaluation may cause teachers to increase the effort they allocate to learning new skills. For example, teachers may feel social pressure to improve over time and to take up peer feedback, not just general pressure to score well. Still, even absent evaluation, teachers likely have some motivation to learn new skills (Dixit 2002), although that effort may be inefficiently allocated across different skills.

Skill improvement may also explain improvements for observer teachers even though they were not actually scored. Observers presumably learned new evaluation skills—how to assess the teaching practices of others—and could have used those new skills, to some extent, for self-evaluation. The motivation to self-assess could be intrinsic if teachers are motivated agents (Dixit 2002), or the motivation could be new social pressures. For example, observers might anticipate being asked questions about their own practices during a debrief with a teacher they observed and scored. And similar to observees, observers would have the rubric's practical advice and normative claims; new information, potentially, about how other teachers in the school scored; and other features of treatment. Finally but most simply, observers had the chance to watch someone else teach, perhaps someone more skilled than the observer themselves. An observer might pick up new practical ideas for how to execute teaching tasks or see a level of effort by a colleague that spurs the observer to raise their own effort.

Perhaps simply providing the rubric to teachers—without any actual observations, scoring, or debriefing—would generate positive treatment effects on teacher performance (Milanowski and Henemen 2001). That rubric-alone hypothesis would be consistent with finding similar effects for observers and observees. Other teachers in treatment schools did not benefit as much as observers and observees, which suggests simple access to the rubric likely is not enough.

We cannot test these mechanism hypotheses empirically, at least not at the level of detail discussed in the previous paragraphs. However, we can provide some relevant evidence by testing for heterogeneity of treatment effects across quantiles of the teacher performance distribution. If peer observation improves teachers' skills and consequently improves student achievement, then treatment effects should be larger for lower-skilled teachers. The estimates shown in figure 1 test this prediction with limitations, and we do find larger effects for lower-performing teachers.[18]

[18] Another test would be to measure teachers' test score contributions in years after the experiment ended. Improvements that persist after the intervention ended would be consistent with skill improvements. However, it may be that the gains we found here would require peer observation continually. Unfortunately, we do not currently have access to the data for school years after the experiment.

In figure 1 the solid line plots QTE estimates, where the outcome variable is teacher value-added scores in the same units as all the other outcomes. The dotted lines mark cluster (school) bootstrap 95% confidence intervals. Value added is shorthand for a teacher's contribution to student GCSE test scores.

We estimate value-added scores using a student fixed effects approach. Because the lagged dependent variable approach is more common in the literature than student fixed effects, we highlight two notes about interpretation of student fixed effects value-added scores. First, the student fixed effects approach tends to understate differences between teachers relative to the lagged dependent variable (Kane and Staiger 2008). This may partly explain the relatively small variance of the QTE in figure 1. Second, in our setting treatment is constant within student (and within teacher); thus, our value-added scores

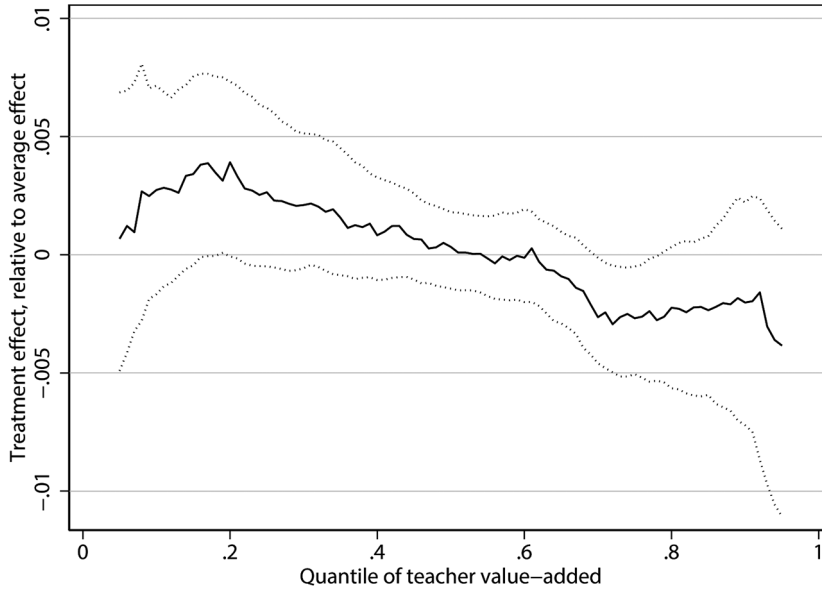

FIG. 1.—Quantile treatment effect estimates. The solid line plots unconditional quantile treatment effects, estimated using recentered influence function regressions with teacher observations. Regressions include randomization block fixed effects. The dotted lines mark the cluster (school) bootstrap 95% confidence intervals (1,000 iterations). The dependent variable is teacher value-added scores. We estimate value added for teachers using the methods described in Kane and Staiger (2008) and Chetty, Friedman, and Rockoff (2014a). Our data include only two school years, so we adopt the assumption that a given teacher's value added is fixed over those 2 years. The student test score measure is math or English General Certificate of Secondary Education tests taken at the end of year 11, and we use student fixed effects in the residualization step.

and QTE estimates are net of the average treatment effect. The zero line in figure 1 represents the average treatment effect, and the QTEs are relative to the 0.073σ average effect.[19]

The pattern in figure 1 does suggest, with limitations, that teacher performance gains (treatment effects) were larger among otherwise low-performing teachers, as would be consistent with the learning new skills mechanism. The QTE point estimates slope downward, with the largest gains near the 25th percentile, although these estimates are somewhat noisy. At or near the 75th percentile the effects are statistically significantly less than zero (i.e., effects were smaller than the average treatment effect); effects are marginally significant near the 20th percentile. The prediction we are testing here is about lower- versus higher-performing teachers. Compare teachers in the 15th–35th percentile range with those in the 65th–85th percentile range. In this comparison we can say that the lower-performing group saw larger treatment effects than the higher-performing group, and the differences are statistically significant.

Interpreting our QTE estimates as effects on low- or high-performing teachers assumes that treatment did not change any teacher's status as low or high performing. The strictest form of this assumption is rank invariance: that treatment did not affect a teacher's performance rank even if it affected her performance level. Some violation of strict rank invariance would not necessarily threaten conclusions about low- and high-performing groups of teachers. However, even without this further assumption, the estimates in figure 1 can be interpreted as treatment effects on the distribution of teacher performance. In that interpretation, our estimates suggest that treatment reduced the variation in teacher performance by improving the relative performance of the lower half of the distribution. Low-performing teachers in treatment schools were still low performing, but they nevertheless outperformed low-performing teachers in control schools.

Two additional notes on mechanisms: First, while our outcome is student test scores, it is unlikely that the treatment mechanism operated through students independent of their teachers. The GCSE exams are quite high

[19] We cannot do the lagged dependent variable approach. Our prior achievement measures are Key Stage 2 exams scores, which occur at year $(t - 5)$. A value-added model using $Y_{im,t-5}$ instead of $Y_{im,t-1}$ would require, implausibly, that student assignments to teachers in year t do not depend on inputs or performance between $(t - 5)$ and $(t - 1)$. In practice, if we re-create fig. 1 using this $Y_{im,t-5}$ approach, the pattern of QTE is similar but noisier. Interpreting any value-added score as the causal effect of teachers on student achievement requires a conditional independence assumption, specifically, the conditional independence of teacher assignments to students. Rothstein (2010) points out that this assumption is quite strong, likely implausible, when the student fixed effects are estimated using data over time for a given student; e.g., fifth-grade math teacher assignment cannot depend on new information from fourth grade. However, in our case the two assignments are year 11 English teacher and year 11 math teacher; both are made at the same time with access to the same information.

stakes for students, but equally so for both treatment and control students. Perhaps treatment teachers encouraged or otherwise elicited more effort from their students. But eliciting student effort is one key component of teacher performance; teacher effects on student effort are always a mechanism in estimates of test score value added.

Second, in this experiment and similar papers, the choice of who should be the observer (mentor) and who the observee (mentee) is a key feature of the treatment bundle and is thus potentially relevant to mechanisms. Our "randomly assign roles" design is quite different from the conventional design, where roles are determined by prior performance or experience. Still, despite the random assignment design, our observer teachers may have been higher performing or more experienced. The differences in table 2 suggest that observers had higher-achieving students, and often more experienced or higher-performing teachers are assigned to higher-achieving students. This possibility does not threaten the identification of the average treatment effect, 0.073σ, but would be relevant to its interpretation, just as it would be in all similar (quasi) experiments.

VI. Costs and Returns

The primary costs of the peer evaluation program were teacher time and effort. Our conservatively high estimate is that each observation required 4 hours of teacher labor, or about £100 for each observation (£50 per teacher). However, teachers were not given extra pay or other compensation for participating in the experiment. Thus, presumably, participation came at some opportunity cost of other job tasks neglected or reduced leisure, which may not be fully captured by the £100 estimate. Adding in the relatively fixed costs for equipment and initial training, the average total cost was just under £450 per teacher per year, or about 1.1% of the average teacher's salary. Complete details of our cost estimates and other estimates in this section are provided in appendix D.

With these relatively small costs, the program compares favorably in cost-effectiveness terms to other educational interventions. First, the formal peer evaluation program studied in Taylor and Tyler (2012) had a similar treatment effect but cost $7,500 per teacher, with the higher costs due mostly to employing specialized, highly trained former teachers as evaluators. Second, in the Project STAR experiment, reducing class size by 30% improved test scores by 0.15σ–0.19σ (Schanzenbach 2006). Those class size gains are more than double the peer observation 0.073σ effect, but reducing class size by 30% requires a 30% increase in labor costs compared with perhaps 1.1% for peer observation. Last, our average effect of 0.073σ is similar to the gain from adding 2–4 weeks of additional class time to the school year (Sims 2008; Fitzpatrick, Grissmer, and Hastedt 2011; Aucejo and Romano 2016), but the extra weeks would presumably require a 5%–10% increase in labor costs.

A cost-benefit analysis requires projecting students' future earnings increases based on test scores, which is always challenging. Still, students' GCSE exam scores—taken at age 16—do influence college going and do predict future earnings (Mcintosh 2006; Hayward, Hunt, and Lord 2014).[20] Our back-of-the-envelope estimates, described in appendix D, suggest that the future earnings gains could plausibly be more than an order of magnitude larger than the program costs.

VII. Conclusion

In this paper we report improvements in teachers' job performance, as measured by their contributions to student test scores, resulting from a program of low-stakes peer evaluation. In randomly assigned treatment schools, teachers visited the classrooms of other teachers in the school and scored the teaching they observed using a structured rubric. Students in treatment schools scored 0.07σ higher on high-stakes GCSE exams in math and English (ITT). In treatment schools that took up the program, students scored at least 0.09σ higher (low-bound TOT). Explanations for the effects of this peer evaluation cannot be based on explicit incentive structures typical in other formal evaluation settings, as these were absent by design; rather, the effects likely operate through teachers gaining new information about their own or others' performance and subsequently improving their teaching skills.

Our paper contributes most directly to the literature on how evaluation affects teacher performance. While most of that literature focuses on the effect of incentives—monetary bonuses, the threat of dismissal—linked to evaluation measures, our contribution is to focus on the effect of the evaluation measures themselves. This paper's experimental comparison is measures versus no measures, with no (explicit) incentives attached to the measures. The roughly 0.07σ average improvement we find is much larger than the improvement in Rockoff et al. (2012), perhaps in part because the current experiment focused on measures of teaching inputs while the Rockoff et al. experiment focused on a measure of output. Other (quasi) experiments have focused on similar observation-rubric input-based measures and found similar benefits (Taylor and Tyler 2012; Steinberg and Sartain 2015; Briole and Maurin 2019), but those previous studies examined formal evaluation programs with explicit incentives attached to the scores.

Two additional results are notable. First, we find no additional benefit from doubling the number of observations—from 1.6 to 2.9 per observee per year. Ours is the first study to directly and experimentally investigate this key feature of evaluation design. Second, we find that observer teachers benefited just as much as observee teachers, and they perhaps benefited more.

[20] Although from different contexts, the evidence in Chetty, Friedman, and Rockoff (2014b) and Deming et al. (2016) lends credibility to arguments that link teacher-caused (school-caused) student test score gains to future income gains.

This paper is the first to randomly assign teacher roles in such "evaluator-evaluatee" or "mentor-mentee" relationships. In our experiment, the observer was often (about half the time) a less experienced and lower-skilled teacher than the observee teacher. Our results question the conventional wisdom that observers and mentors must be selected for a history of high performance. Moreover, the fact that observers in our experiment also benefited themselves changes the cost-benefit calculus of such a program. One potential implication is quite striking: schools that outsource the evaluator role and schools in which observation/evaluation are just another task for school leaders are missing out on more than half of the potential gain to observation.

The benefits of peer evaluation documented in this experiment suggest practical and promising policy ideas for improving the job performance of a sizable workforce. Relative to other educational interventions, this program seems both politically and financially attractive. Politically, the program is likely less threatening because of the absence of strong or explicit consequences attached to the scores and because observers are peers rather than outside experts or school leaders. Financially, it is likely less threatening because it is an inexpensive intervention, at least in budget terms, with relatively large potential returns.

References

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics* 131, no. 3:1415–53.

Aucejo, Esteban M., and Teresa Foy Romano. 2016. Assessing the effect of school days and absences on test score performance. *Economics of Education Review* 55:70–87.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. 2019. An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review* 73:101919.

Briole, Simon, and Eric Maurin. 2019. Does evaluating teachers make a difference? IZA Discussion Paper no. 12307, IZA Institute of Labor Economics, Bonn, Germany.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104, no. 9:2593–632.

———. 2014b. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104, no. 9:2633–79.

Danielson, Charlotte. 2007. *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Dee, Thomas S., and James Wyckoff. 2015. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34, no. 2:267–97.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics* 98, no. 5:848–62.

Dixit, Avinash. 2002. Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources* 37, no. 4:696–727.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. Unconditional quantile regressions. *Econometrica* 77, no. 3:953–73.

Fitzpatrick, Maria D., David Grissmer, and Sarah Hastedt. 2011. What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review* 30, no. 2:269–79.

Garet, Michael S., Andrew J. Wayne, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske. 2017. *The impact of providing performance feedback to teachers and principals*. Publication no. 2018-4001. Washington, DC: Institute for Education Sciences, US Department of Education.

Garrett, Rachel, and Matthew P. Steinberg. 2015. Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis* 37, no. 2:224–42.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *Identifying effective teachers using performance on the job*. Hamilton Project Policy Brief no. 2006-01. Washington, DC: Brookings Institution.

Hanushek, Eric A., and Steven G. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review* 100, no. 2:267–71.

Hayward, Hugh, Emily Hunt, and Anthony Lord. 2014. *The economic value of key intermediate qualifications: Estimating the returns and lifetime productivity gains to GCSEs, A levels and apprenticeships*. Department for Education Report no. DFE-RR398A. London: Department for Education.

Jackson, Kirabo, and Alexey Makarin. 2018. Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment. *American Economic Journal: Economic Policy* 10, no. 3:226–54.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. Teacher effects and teacher-related policies. *Annual Review of Economics* 6, no. 1:801–25.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle: Bill and Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper no. 14607, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 46, no. 3:587–613.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research* 88, no. 4:547–88.

McIntosh, Steven. 2006. Further analysis of the returns to academic and vocational qualifications. *Oxford Bulletin of Economics and Statistics* 68, no. 2:225–51.

Milanowski, Anthony T., and Herbert G. Heneman. 2001. Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education* 15, no. 3:193–212.

Murphy, Richard, Felix Weinhardt, and Gillian Wyness. 2021. Who teaches the teacher? A RCT of peer-to-peer observation and feedback in 181 schools. *Economics of Education Review* 82:102091.

Neal, Derek. 2011. The design of performance pay in education. In *Handbook of the economics of education*, vol. 4, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 495–550. Amsterdam: Elsevier.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary E. Laski. 2020. Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy* 12, no. 1:359–88.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review* 102, no. 7:3184–213.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125, no. 1:175–214.

———. 2015. Teacher quality policy when supply matters. *American Economic Review* 105, no. 1:100–130.

Schanzenbach, Diane Whitmore. 2006. What have researchers learned from Project STAR? *Brookings Papers on Education Policy* 9:205–28.

Sims, David P. 2008. Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review* 27, no. 1:58–68.

Slater, Helen, Neil M. Davies, and Simon Burgess. 2012. Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics* 74, no. 5:629–45.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. Searching for effective teachers with imperfect information. *Journal of Economic Perspectives* 24, no. 3:97–118.

Steinberg, Matthew P., and Lauren Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy* 10, no. 4: 535–72.

Taylor, Eric S., and John H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102, no. 7:3628–51.

Worth, Jack, Juliet Sizmur, Matthew Walker, Sally Bradshaw, and Ben Styles. 2017. *Teacher observation: Evaluation report and executive summary*. London: Education Endowment Foundation.