Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools

Appendix

Simon Burgess, University of Bristol
Shenila Rawal, Oxford Partnership for Education Research and Analysis
Eric S. Taylor, Harvard University

September 2020

A. Additional Tables and Figures

Appendix Table A1—Teacher role experiment baseline covariate summary measure

	Year one	Year two
	(1)	(2)
Teacher role (relative to Observer)		
Observee	-0.047+	-0.094
	(0.029)	(0.059)
Both roles	-0.009	0.017
	(0.028)	(0.058)

Note: Each column reports results from a separate least squares regression. The dependent variable is a summary index of pre-treatment covariates, constructed as follows: Using only control school observations, we regress outcome test score (student math or English GCSE score in student standard deviation units) on all pre-treatment covariates; and then use the estimated coefficients calculate fitted values for the treatment schools. This fitted value is the summary index. In the regression reported in this table, we regress the summary index on the teacher role indicators and department fixed effects. Heteroskedasticity-cluster robust standard errors in parentheses, with clustering at the teacher level.

+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A2—First-stage results for Table 3

	Dep. var. = endogenous treatment in Table 3…		
	Col (3)	Col (4)	Col (5)
School randomly assigned to treatment	0.787**	0.551**	0.469**
	(0.062)	(0.068)	(0.076)
Adjusted *R*-squared	0.661	0.446	0.346
Observations	56,148	56,148	56,148

Note: Each column reports results from a separate least squares regression, each a first-stage regression associated with the 2SLS estimates in Table 3. Heteroskedasticity-cluster robust standard errors in parentheses, clustering at the school level.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A3—Treatment effect estimates for subsamples

	Subsample			
	Math scores	English scores	Grade 11 in 2014-15	Grade 11 in 2015-16
	(1)	(2)	(3)	(4)
School randomly assigned to treatment	0.044	0.102*	0.106**	0.037
	(0.031)	(0.040)	(0.036)	(0.034)
Adjusted *R*-squared	0.398	0.332	0.296	0.427
Observations	28,074	28,074	28,410	27,738

Note: Each column reports results from a separate least squares regression, with student-by-subject observations. Estimation is identical to Table 3 column 2, except that the estimation sample is limited to a subsample described in the column header. Heteroskedasticity-cluster robust standard errors in parentheses, clustering at the school level.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A4—Dose condition effects on number of observations completed

	(1)
Treatment school	1.603**
	(0.254)
High dose department	1.334**
	(0.310)
Adjusted *R*-squared	0.432
Observations	56,148

Note: Results from a least squares regression, with student-by-subject observations. The dependent variable is the number of observations completed per observee teacher; specifically, (i) the number of observations completed by the school before the end of the year in which the student took the GCSE exam, divided by (ii) number of observee teachers. The specification also includes an indicator for math observation, an indicator for cohort 1, and the pre-treatment covariates listed in Table 1. When a pre-treatment covariate is missing, we replace it with zero and include an indicator variable = 1 for missing on the given characteristic. Heteroskedasticity-cluster robust standard errors in parentheses, clustering at the school level.

+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A5—Effects by teacher role, additional results

		Year				Year	Year
	Pooled	two	Pooled			one	two
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher role (relative to Observer)							
Observee	-0.080*	-0.061	-0.029	-0.042	-0.059*	-0.046	-0.005
	(0.041)	(0.057)	(0.029)	(0.029)	(0.029)	(0.033)	(0.047)
Both roles	-0.038	-0.001	-0.023	-0.023	-0.048	-0.048	0.013
	(0.041)	(0.058)	(0.029)	(0.030)	(0.029)	(0.034)	(0.046)
Pre-treatment covariates	√	√					
Student fixed effects			√	√	√	√	√
Year one teacher controls				√			
Year one teacher FE	√	√			√		
Adjusted R-squared	0.419	0.585	0.674	0.674	0.695	0.676	0.672
Observations	15,077	6,390	15,077	15,077	15,077	8,687	6,390

Note: Each column reports results from a separate least squares regression, with student-by-subject observations. See Table 5 notes for details on estimation. Heteroskedasticity-cluster robust standard errors in parentheses, with clustering at the teacher level.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A6—Robustness to degree of imbalance

	Low imbalance	High imbalance
	(1)	(2)
Teacher role (relative to Observer)		
Observee	-0.057	-0.188**
	(0.094)	(0.084)
Both roles	-0.054	0.017
	(0.093)	(0.078)
Adjusted *R*-squared	0.070	0.072
Observations	6,171	8,906

Note: Each column reports results from a separate least squares regression, with student-by-subject observations. The estimation details are identical to Table 5 column 1, except that here each column is estimated using a subsample. The two subsamples relatively "low imbalance" and "high imbalance" are defined in the following way: First, for each student, we convert the available pre-treatment covariates into a scalar index measure. Using the control sample, we regress GCSE score on those covariates, and then calculate the fitted GCSE score for treatment cases. That fitted score is our index. Second, for each treatment school, we estimate the mean difference in that index between teachers randomly assigned to be observers and observees as in Appendix Table A1. We define relatively "low imbalance" schools as schools where the absolute fitted-score difference of below the median (roughly 0.12σ), and "high imbalance" schools above the median. Heteroskedasticity-cluster robust standard errors in parentheses, with clustering at the teacher level. + indicates p<0.10, * 0.05, and ** 0.01

Appendix B. Experiment Design and Setting, Additional Details

B.1 Rational for and Development of the Experiment

Funding for the experiment was provided by the Education Endowment Foundation (EEF). In 2013 we approached the EEF with a proposal to fund a project to test an intervention to raise teacher effectiveness. From the start our proposal was to experimentally test whether peer observation improves teacher job performance. Our rational for this proposal had three parts:

Why study ways to improve teacher job performance?—First, differences in teacher job performance are large, with lasting impacts on students. When performance is measured by a teacher's contribution to student achievement test scores, the difference between a bottom quartile and top quartile teacher is often 10-25 percent of the total variation in student scores (see reviews in Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014). While many of these "teacher value-added" estimates come from U.S. elementary and middle schools, similar estimates have been documented elsewhere, including England (Slater, Davies, and Burgess 2011). Chetty, Friedman, and Rockoff (2014b) further suggests that teachers who cause larger test score gains, also cause improvements in adult outcomes years later, e.g., more likely to go to college, earn higher salaries, less likely to have a teenage pregnancy. Advances in measurement, especially in the last two decades, have made these teacher performance differences more salient to researchers, policymakers, and school managers.

Second, those differences in teacher performance suggest there may be opportunities to meaningfully improve student outcomes through better management of the teacher workforce, for example, through strategies for teacher selection, training, incentives, etc. Yet, empirical evidence documenting successful strategies remains scarce (Jackson, Rockoff, and Staiger 2014). The most frequently proposed strategy, at least among economists, is probationary screening—measure on-the-job performance early and dismiss observed low performers—which seemingly only requires good measures of performance (Gordon, Kane, and Staiger 2006). However, the benefits in equilibrium likely depends more on how labor supply, and thus labor costs, respond (Staiger and Rockoff 2010, Rothstein 2015). Pay for performance, and similar incentive-based approaches, are another commonly proposed strategy, but (quasi-)experimental tests have returned mixed results at best (Neal 2011, Jackson, Rockoff, and Staiger 2014). Finally, schools' traditional strategy is formal training courses for current teachers, but, despite the enormous amount spent annually on such "professional development" courses, there is effectively no (quasi-)experimental evidence that they do improve teaching (Yoon et al. 2007, Jackson, Rockoff, and Staiger 2014).

Why focus on a peer observation strategy?—We believed "peer observation" was a promising strategy for three reasons. First, there was new (at that time) quasi-experimental evidence of positive effects. Studying teachers in Cincinnati, Ohio, Taylor and Tyler (2012) found that teacher performance improved by roughly 0.10σ as a result of peer evaluation; the outcome measure was student test scores, but the evaluation program used rubric-scored classroom observations.[1] Second, the Cincinnati results were consistent with a hypothesis that the process of evaluation might cause teachers to improve their job skills. Specifically, the Cincinnati teachers' performance remained higher in the years after they were no longer being evaluated and had no evaluation incentives. This "evaluation as professional development" idea was, and still is, widely advocated in the education sector, in part to justify spending on evaluation. Third, peer observation was likely to be relatively inexpensive, compared to the other typical strategies for improving teacher performance, e.g., monetary bonuses or the costs of turnover generated by probationary screening. For a discussion of costs see Section 5 in the main paper and Appendix D.

[1] Additionally, there were also early positive results from a program in Chicago, where school principals conducted the rubric-scored classroom observations (Steinberg and Sartain 2015).

Why a new experiment?—First, while encouraging, the Taylor and Tyler (2012) result was one result, and a quasi-experimental result. The strategy of "peer observation" had received almost no empirical attention, especially in comparison to, for example, "pay for performance." There were plenty of open questions about generalizability, mechanisms, etc. As examples, our proposed experiment occurred in a different country, at a different grade level, and with a sample of teachers more diverse in their teaching experience. The Cincinnati observations were part of a long-running district program, but our observation program was new to participating schools. Second, the experimental approach allowed us to test specific features of the peer observation design. In the end, given sample and power budget constraints, we settled on experimentally varying the number of observations and teachers' roles. These additional randomizations would, we expected, generate distinctively new contributions.

B.2 Additional Details

Key details of the experiment and setting are described in Section 1 of the main paper. Here we provide some additional details as a complement to Section 1.

B.2.1 Grade Level and Subjects, Outcome Measures

Our choice of grade level and subjects was, in effect, a choice of student test score outcome measure. Student scores as outcome measure is a fundamental feature of our study design, in part because it allows us to keep separate (i) the intervention measure of teacher performance, rubric-scored observations from (ii) the outcome measure of teacher performance, contributions to student scores.

Our choice set was effectively two options. We chose the General Certificate of Secondary Education (GCSE) exams, which students take at the end of year 11 when they are typically age 16. The alternative option was the Key Stage 2 exams taken at year 6, age 11.[2] One important advantage of the GCSE outcome

[2] Routine student testing in England occurs at four grade levels. In addition to the "GCSEs" and "Key Stage 2" exams, there is also: A "Key Stage 1" exam in year 2/age 7, but these exams are

is that we have a pre-experiment test score measure (i.e., the Key Stage 2 scores) to improve precision. A different approach would have been to split our resources across the two options: GCSE and Key Stage 2. However, dividing a fixed budget in two, we would not have had sufficient statistical power to detect heterogeneity in the treatment effect for GCSEs compared to the effect Key Stage 2.

The GCSE exams, especially math and English, are relatively high stakes for students. Many future employment and training opportunities are based on achieving minimum GCSE scores. The exams can partly inform college/university admissions, but students bound for college still have two more years of secondary schooling after the GCSE exams. At age 18 students remaining in school take the "A-Level" exams. The relatively high stakes nature of the GCSEs is an advantage. First, students already have strong incentives to do well, and motivated teachers thus have incentives to contribute. This reduces the scope for transient increase (decrease) in effort prompted by the experiment *per se*. Second, GCSE scores do have a more direct link to future education and labor market outcomes—more direct than tests at age 11 (Mcintosh 2006, Hayward, Hunt, and Lord 2014). This strengthens the economic and educational significance of any positive treatment effects.

B.2.2 Recruiting Schools

We contacted 1,097 schools and invited each to participate in the experiment, with the goal of recruiting 120. The 1,097 were, at least at the time, nearly all high-poverty public (state) secondary schools in England. "High-poverty" is defined as being in the top half of all schools measured by the percent of students qualifying for free school meals. "Nearly all" because we excluded boarding schools and single-gender schools, as well as schools where the study

graded by the students' teacher, which would have clear complications for our experiment. And the "A-Level" exams at the end of secondary school/age 18.

funder EEF was conducting different interventions (i.e., Lancashire, Merseyside, and Somerset). Student achievement (test scores) were not used as a criterion for inviting schools.

The headteacher (principal) at each of the 1,097 schools was initially contacted by letter in the summer of 2014. The recruitment materials sent to headteachers described the peer observation program expectations in detail, including: observations would be structured and scored using a well-established observation rubric, the collection of scores would be via tablet computers, the number and frequency of observations, the fact that only some teachers would be observees while others would be the observers (as opposed to outside observers), an overview of the initial training, etc. The materials also reiterated the random assignment nature of the study, and thus only half of schools would ultimately be asked to implement the peer observation program.

Of the invited schools, 92 initially volunteered to participate (8.5 percent). We did not collect systematic data on why schools (headteachers) were motivated to volunteer. One minor motivation may have been the small incentive of £1,000 per participating school, which was small relative to a typical secondary school budget of over £3m. Schools were also told that, if they ended up in the treatment condition, they could keep the iPad computers. Some may have been motivated by the desire to support research. A likely motivation for all schools was an expectation that the peer observations could benefit teachers and students. The recruitment materials did not make any strong claims about expected benefits, nor describe any prior research, but the implication in any study is that the researchers and funders expect positive benefits. Anecdotal information from the recruiters and trainers is consistent with schools' expecting benefits as a motivation. Some schools emphasized the peer-to-peer feature of the observations as novel and more conducive to teacher development than observations by administrators. Other schools mentioned the formal rubric and its history and evidence base.

Before random assignment, the 92 schools were asked to confirm their participation by signing a Memorandum of Understanding, and by providing class rosters (i.e., NPD student IDs linked to a study-specific teacher ID). Ten schools dropped out before randomization. We randomized the remaining 82 schools to treatment and control.

B.2.3 Observation Scoring Rubric

Observer teachers scored their observations of observee peers using a rubric called the *Framework for Teaching* (Danielson 2007, "FFT"). The FFT rubric has been used by schools for more than two decades, and has been increasingly used in research (for example, Kane et al. 2011, Kane et al. 2013, Bacher-Hicks et al. 2017). FFT is the rubric used by peer evaluators in the Taylor and Tyler (2012) setting. In our experiment teachers were scored on 10 rubric items, known as "standards" grouped in two "domains." The full rubric is shown in Figure B1 below.

Conventionally each FFT item is scored on a 1-4 integer scale, with 1-4 corresponding to "Ineffective" through "Highly Effective". We instead asked observers to use a 1-12 integer scale. "Highly Effective" could be 10, 11, or 12 roughly corresponding to "Highly Effective +", "Highly Effective", and "Highly Effective –"; with similarly 3 score values for "Effective", "Basic", and "Ineffective" as well. The 1-12 scale was motivated in part by the typical skewness toward scores of 3-4 on evaluation ratings using a 1-4 scale generally, and with teacher observation rubrics specifically (Kane et al. 2011, Kraft and Gilmour 2016). This tendency was confirmed in our pilot stage work with schools. A second motivation for the 1-12 scale was to avoid confusion with a similar 1-4 scale used by Ofstead.

Pooling the 10 standards, the average rubric item score was 9.05 with a standard deviation of 2.10. If we convert the 1-12 scale to the more common 1-4 scale (i.e., 12 = 4.33, 11 = 4, 10 = 3.66, 9 = 3, and so on) the average item score is 3.35 (st.dev. 0.70). This mean is quite similar to other contexts, for example, Kane et al. (2011) where the mean is 3.23. However, in our setting the standard deviation is a wider 0.70 versus, for example, 0.49 in Kane et al. (2011). The wider variation may be due in part to the 1-12 scale.

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

Observers recorded data and scores using an iPad tablet computer and software provided by RANDA Solutions. In the iPad app, observers could access the complete FFT rubric, record scores, and make notes during their observations. The centrally-stored database of observations allowed the research team to monitor progress of individual schools, and contact those who were clearly lagging.

B.2.4 Training for Teachers

Treatment teachers were provided initial training on the FFT rubric and other aspects of the peer observation program. Though, to be clear, this training was not as extensive as administrators or other evaluators often receive in formal evaluation systems, nor as extensive as the training of raters in research studies that use the same observation rubrics. First, one "lead" teacher from each school participated in the project's main training event. These lead teachers were the contact and coordinator for the program in each school. Lead teachers also provided training to the other teachers in their own schools. Second, the research team conducted six additional regional training sessions for teachers. The regional organization was designed to minimize travel time. The research team also visited five schools individually for training because the five did not fit into one of the six regions. Some follow-up training and support occurred by phone.

The content of the training focused on the use of the rubric to evaluate observed teaching, including discussing the conceptual framework, comparing FFT to other frameworks, practicing applying the rubric, and answering questions from trainees. The training also included an introduction to the research project and experimental design, and time to learn the RANDA software and establish log-in accounts.

B.2.5 Teacher Data

We did not collect any data on individual teachers, with the one exception of class rosters which listed the students assigned to each teacher. Additionally, while the class rosters used student IDs that could be linked back to NPD data, the teacher IDs were randomly assigned numbers only created for and used in the research project and thus could not be linked to teachers' names or other characteristics. The rational for this approach, acknowledging the limitations it creates, is that it would protect individual teachers' confidentiality and thus promote greater teacher participation. Individual teachers could not be identified by anyone outside of their own school.

References

In addition to those listed in the main paper's references list:

Kraft, M. A., & Gilmour, A. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753.

Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement: Issues & Answers Report*, REL 2007–No. 033. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

DOMAIN 1: THE CLASSROOM ENVIRONMENT				
Component	**Ineffective (1-3)**	**Basic (4-6)**	**Effective (7-9)**	**Highly Effective (10-12)**
1a Creating an Environment of Respect and Rapport	Classroom interactions, both between the teacher and students and among students, are negative, inappropriate, or insensitive to students' cultural backgrounds, ages and developmental levels. Student interactions are characterised by sarcasm, put-downs, or conflict.	Classroom interactions, both between the teacher and students and among students, are generally appropriate and free from conflict, but may reflect occasional displays of insensitivity or lack of responsiveness to cultural or developmental differences among students.	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.	Classroom interactions, both between teacher and students and among students, are highly respectful, reflecting genuine warmth and caring and sensitivity to students' cultures and levels of development. Students themselves ensure high levels of civility among members of the class.
1b Establishing a Culture for Learning	The classroom environment conveys a negative culture for learning, characterised by low teacher commitment to the subject, low expectations for student achievement, and little or no student pride in work.	The teacher's attempts to create a culture for learning are partially successful, with little teacher commitment to the subject, modest expectations for student achievement, and little student pride in work. Both teacher and students appear to be only "going through the motions."	The classroom culture is characterised by high expectations for most students and genuine commitment to the subject by both teacher and students, with teacher demonstrating enthusiasm for the content and students demonstrating pride in their work.	High levels of student energy and teacher passion for the subject create a culture for learning in which everyone shares a belief in the importance of the subject and all students hold themselves to high standards of performance they have internalized.

DOMAIN 1: THE CLASSROOM ENVIRONMENT (cont.)				
Component	**Ineffective (1-3)**	**Basic (4-6)**	**Effective (7-9)**	**Highly Effective (10-12)**
1c Managing Classroom Procedures	Much teaching time is lost because of inefficient classroom routines and procedures for transitions, handling of supplies, and performance of non-teaching duties. Students not working with the teacher are not productively engaged in learning. Little evidence that students know or follow established routines.	Some teaching time is lost because classroom routines and procedures for transitions, handling of supplies, and performance of non-teaching duties are only partially effective. Students in some groups are productively engaged while unsupervised by the teacher.	Little teaching time is lost because of classroom routines and procedures for transitions, handling of supplies, and performance of non-teaching duties, which occur smoothly. Group work is well-organised and most students are productively engaged while working unsupervised.	Teaching time is maximised due to seamless and efficient classroom routines and procedures. Students contribute to the seamless operation of classroom routines and procedures for transitions, handling of supplies, and performance of non-instructional duties. Students in groups assume responsibility for productivity.
1d Managing Student Behaviour	There is no evidence that standards of conduct have been established, and there is little or no teacher monitoring of student behaviour. Response to student misbehaviour is repressive or disrespectful of student dignity.	It appears that the teacher has made an effort to establish standards of conduct for students. The teacher tries, with uneven results, to monitor student behaviour and respond to student misbehaviour.	Standards of conduct appear to be clear to students, and the teacher monitors student behaviour against those standards. The teacher response to student misbehaviour is consistent, proportionate, appropriate and respects the students' dignity.	Standards of conduct are clear, with evidence of student participation in setting them. The teacher's monitoring of student behaviour is subtle and preventive, and the teacher's response to student misbehaviour is sensitive to individual student needs and respects students' dignity. Students take an active role in monitoring the standards of behaviour.

DOMAIN 1: THE CLASSROOM ENVIRONMENT (cont.)

Component	Ineffective (1-3)	Basic (4-6)	Effective (7-9)	Highly Effective (10-12)
1e Organising Physical Space	The physical environment is unsafe, or some students don’t have access to learning. There is poor alignment between the physical arrangement of furniture and resources and the lesson activities.	The classroom is safe, and essential learning is accessible to most students; the teacher’s use of physical resources, including computer technology, is moderately effective. The teacher may attempt to modify the physical arrangement to suit learning activities, with limited effectiveness.	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.	The classroom is safe, and the physical environment ensures the learning of all students, including those with special needs. Students contribute to the use or adaptation of the physical environment to advance learning. Technology is used skilfully, as appropriate to the lesson.

DOMAIN 2: TEACHING				
Component	**Ineffective (1-3)**	**Basic (4-6)**	**Effective (7-9)**	**Highly Effective (10-12)**
2a Communicating with Students	Expectations for learning, directions and procedures, and explanations of content are unclear or confusing to students. The teacher's written or spoken language contains errors or is inappropriate for students' cultures or levels of development.	Expectations for learning, directions and procedures, and explanations of content are clarified after initial confusion; the teacher's written or spoken language is correct but may not be completely appropriate for students' cultures or levels of development.	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.	Expectations for learning, directions and procedures, and explanations of content are clear to students. The teacher links the instructional purpose of the lesson to the wider curriculum. The teacher's oral and written communication is clear and expressive, appropriate to students' cultures and levels of development, and anticipates possible student misconceptions. The teacher's explanation of content is thorough and clear, developing conceptual understanding through clear scaffolding and connecting with students' interests. Students contribute to extending the content by explaining concepts to their peers and suggesting strategies that might be used.

DOMAIN 2: TEACHING (cont.)

Component	Ineffective (1-3)	Basic (4-6)	Effective (7-9)	Highly Effective (10-12)
2b Using Questioning and Discussion Techniques	The teacher's questions are of low cognitive challenge or inappropriate, eliciting limited student participation, and recitation rather than discussion. A few students dominate the discussion.	Some of the teacher's questions elicit a thoughtful response, but most are low-level, posed in rapid succession. The teacher's attempts to engage all students in the discussion are only partially successful.	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.	Questions reflect high expectations and are culturally and developmentally appropriate. Students formulate many of the high-level questions and ensure that all voices are heard.
2c Engaging Students in Learning	Activities and assignments, materials, and groupings of students are inappropriate for the learning outcomes or students' cultures or levels of understanding, resulting in little intellectual engagement. The lesson has no clearly defined structure or is poorly paced.	Activities and assignments, materials, and groupings of students are partially appropriate for the learning outcomes or students' cultures or levels of understanding, resulting in moderate intellectual engagement. The lesson has a recognisable structure but is not fully maintained and is marked by inconsistent pacing.	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.	Students, throughout the lesson, are highly intellectually engaged in significant learning and make material contributions to the activities, student groupings, and materials. The lesson is adapted as needed to the needs of individuals, and the structure and pacing allow for student reflection and closure.

DOMAIN 2: TEACHING (cont.)

Component	**Ineffective (1-3)**	**Basic (4-6)**	**Effective (7-9)**	**Highly Effective (10-12)**
2d Use of Assessment	Assessment is not used in teaching, either through monitoring of progress by the teacher or students, or adequate feedback to students. Students are not aware of the assessment criteria used to evaluate their work, nor do they engage in self- or peer-assessment. .	Assessment is occasionally used in teaching, through some monitoring of progress of learning by the teacher and/or students. Feedback to students is uneven, and students are aware of only some of the assessment criteria used to evaluate their work. Students occasionally assess their own or their peers' work.	Assessment is regularly used in teaching, through self- 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.	Assessment is used in a sophisticated manner in teaching, through student involvement in establishing the assessment criteria, self-or peer assessment by students, monitoring of progress by both students and the teacher, and high-quality feedback to students from a variety of sources. Students use self-assessment and monitoring to direct their own learning.
2e Demonstrating Flexibility and Responsiveness	The teacher adheres to the lesson plan, even when a change would improve the lesson or address students' lack of interest. The teacher brushes aside student questions; when students experience difficulty, the teacher blames the students or their home environment.	The teacher attempts to modify the lesson when needed and to respond to student questions, with moderate success. The teacher accepts responsibility for student success but has only a limited repertoire of strategies to draw upon.	The teacher promotes the successful learning of all students, making adjustments as needed to lesson plans and accommodating student questions, needs, and interests.	The teacher seizes an opportunity to enhance learning, building on a spontaneous event or student interests, or successfully adjusts and differentiates instruction to address individual student misunderstandings. The teacher ensures the success of all students by using an extensive repertoire of teaching strategies and soliciting additional resources from the school or community. .

Appendix C. Treatment Effect Levels for Observers and Observees

The average treatment effect level across all teachers is 0.073σ. If 0.073σ was simply a weighted average of effects for observers, observees, and "both roles" teachers then the effect level for observers would be about 0.097σ, for observees 0.040σ, and for "both roles" 0.082σ. These estimates are simply the solutions to

$$0.073 = \frac{1}{3}(\delta^{VER} + \delta^{VEE} + \delta^{BOTH}) \tag{C1}$$

$$-0.057 = \delta^{VEE} - \delta^{VER} \tag{C2}$$

$$-0.015 = \delta^{BOTH} - \delta^{VER} \tag{C3}$$

where δ terms with superscripts are treatment effects for subgroups, and C2 and C3 are taken from Table 5 column 2.

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, "both roles" teachers, and these other non-participating teachers. Specifically, the equation C1 above should be

$$0.073 = (1 - \lambda)\frac{1}{3}(\delta^{VER} + \delta^{VEE} + \delta^{BOTH}) + \lambda\delta^{NP} \tag{C4}$$

where λ is the proportion of non-participating teachers, and δ^{NP} the treatment effect on non-participating teachers. We observe λ directly in the data, but we do not have a readily available estimate for δ^{NP}. Simplifying, we can combine C4 with C2 and C3 to write

$$\delta^{VER} = 0.097 + \lambda(\delta^{P} - \delta^{NP}) \tag{C5}$$

where $\delta^{P} = \frac{1}{3}(\delta^{VER} + \delta^{VEE} + \delta^{BOTH})$.

Even without further estimation, equation C5 makes clear an intuitive conclusion: The effects for observers will be larger than 0.097σ, as long as the treatment effect for non-participants is smaller than the treatment effect for participants, $(\delta^{P} - \delta^{NP}) > 0$. The same applies to observees and "both role" teachers.

We can estimate $(\delta^{P} - \delta^{NP})$, though imperfectly, by adapting our main specification 1. We add two right-hand-side variables: a new indicator variable $P_j = 1$ if teacher j was a participant, and the interaction between P_j and the treatment indicator T_s. Recall that both treatment and control schools provided class rosters before random assignment, and both excluded some teachers from participation at that step. Also, both treatment and controls schools could withdraw consent to use those rosters. Thus we can construct P_j for both treatment and control. Still, selection into $P_j = 1$ or $P_j = 0$ was not randomly assigned, and the selection could differ between treatment and control. For example, through differential assignment of students in year two, or differential withdrawal of consent to use rosters. To address the potential selection we use a student fixed effects approach, identifying $(\delta^{P} - \delta^{NP})$ based on students who had a participating teacher for math and a non-participating teacher for English, or the reverse.

Following this approach, we estimate that the effect for observers was 0.120σ and for observees 0.063σ. The former is simply equation C5 with $\lambda = 0.488$ as observed in the data, and $(\delta^{P} - \delta^{NP}) = 0.047$ from our student fixed effects estimator. The later uses equation C2.

Appendix D. Costs and Returns

The primary inputs to peer evaluation are teacher time and effort. There were the actual classroom observations, of course, typically lasting just 20 minutes. Each observation also presumably required some logistical tasks, plus time for observer and observee to meet and discuss the results. If the observer had to leave her own class of students, another teacher (adult) would need to cover for her. Additionally, each teacher participated in a few hours of initial training.

What did that teacher time and effort cost the schools? As a rough estimate we estimate teacher hours and then multiply by the average teacher wage rate. Our conservatively-high estimate is four hours of teacher labor or £100 for each observation, £50 per participating teacher. Details of all estimates in this section are provided in section D.1 below. However, teachers were not given extra pay or other compensation for participating in the experiment. Thus, presumably, teacher time and effort spent on peer observation came at the cost of other job tasks neglected or reduced leisure. Schools had no extra budgetary costs during the experiment.[1] Still, schools may have had opportunity costs associated with neglected job tasks. On net student achievement rose in math and English, but other unobserved outcomes could have declined. Moreover, in the long run, teachers and schools may not be willing or able to cover the four hours (£100) cost with neglected tasks or leisure. Thus, we include these teacher time costs when comparing cost and returns.

The total cost of the peer evaluation program, we estimate, was just under £450 per teacher per year, or about 1.1 percent of the average teacher's salary. This total cost includes 2.27 observations per observee teacher per year. Most of the £450 cost, however, is the relatively fixed costs of the initial training, tablet

[1] One potential exception is that schools may have spent more on cover (substitute) teacher labor than they otherwise would have, or they may have just reallocated that budget as well.

computers, and software licenses which total £334 per teacher per year. If the peer observation program were to continue beyond just two years, or to expand to other subject departments in the school, these annualized fixed costs per teacher would fall. Again, the cost per observation is likely less than £100.

These costs are not immediately comparable with returns measured in GCSE score gains. Still, we can make a back-of-the-envelope comparison of costs and students' future income gains predicted by GCSE scores. The convention in GCSE-to-income estimates is to compare students who did and did not earn "five good GCSEs," that is, earn a GCSE grade of A*-C in math, English, and at least three other subject areas.[2] Mcintosh (2006) estimates that students who achieve "five good GCSEs" earn a 25-30 percent wage premium over those with lower qualifications; Hayward, Hunt, and Lord (2014) estimate the NPV at £100K in lifetime earnings above what is earned by those who only earn a few A*-C grades. Our focus in this paper is math and English, but earning A*-C in math and English is highly predictive of earning A*-C in three (or more) other subjects.

It is plausible that the treatment pushed many students over the C grade threshold. The average student in our sample scored just below C; the average control student scored one-third of a grade point below C in math and one-quarter point in English. Additionally, we can also estimate treatment effects on the binary outcome = 1 if the student earned an A*-C grade in both math and English. Treatment increased the probability of earning "good GCSEs in math and English" by 3.1 percentage points (*p*-value = 0.085). The experiment's total cost per student was about £50. Together, these estimates suggest the expected cost of moving one additional student into "earned good GCSEs in math and English" was about £1,600

[2] The possible grades are A*, A, B, C, D, E, F, G, and U, where U is failing.

(£50/0.031). And thus also suggest a high rate of return to the intervention (£1.6K costs to £100K in lifetime earnings).[3]

Alternatively, instead of trying to monetize test score gains, we can make cost-effectiveness comparisons with 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; the higher costs are due mostly to employing specialized, highly-trained former teacher as evaluators. Second, in the Project STAR experiment, reducing class size by 30 percent improved test scores by 0.15-0.19σ (Schanzenbach 2006). Those class size gains are more than double the peer observation 0.073σ effect, but reducing class size by 30 percent requires a 30 percent increase in labor costs compared to perhaps 1.5 percent 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 percent increase in labor costs.

D.1 Details on Cost Estimates

Teacher Wage Rate—We use a wage rate of £25 per hour. Annual salaries for qualified teachers in England generally range from £24K starting salary outside of London up to £50K for experienced, high-performing teachers in inner London. Our £25 per hour estimate is £40K divided by 40 weeks per year * 40 hours per week.

Teacher Labor Costs per Observation—We do not have data to measure teacher time use, other than for the classroom observations themselves. Our

[3] Though from different contexts, the evidence in Chetty, Friedman, and Rockoff (2014b) and Deming, Cohodes, Jennings, and Jencks (2016) lends credibility arguments which link teacher-caused (school-caused) student test score gains to future income gains.

estimates here are back-of-the-envelope estimates, and we believe they are conservatively high.

The typical observation lasted 20 minutes. We add an additional 20 minutes for logistical tasks: preparation, time to walk between classrooms, wrap up, etc. Teachers were encouraged to meet together after the observation to discuss the results and strategies for improvement. We add 60 minutes for these post-observation meetings. In total we allocate 100 minutes to each teacher, or 200 minutes total combining observer time and observee time.

However, we view this 200-minutes estimate as conservatively large. For example, during the 20 minutes of observation, the observee is teaching her class just as she otherwise would absent the treatment. Thus, at least those 20 minutes (of the 200) are not necessarily new time costs for the school created by the peer observation. Similarly, meeting to discuss results may have occurred during other existing meetings or replaced other responsibilities.

Finally, we add 40 minutes of time for a third teacher who substituted for the observer in the observer's class during the peer observation. A "cover teacher" was not always required; some observations were done in times when the observee was teaching, but the observer was not. And the "cover teacher" might have been a less-expensive teachers aid, or a more-expensive school administrator.

Our estimate of total teacher labor hours is 240 minutes or 4 hours per observation. Thus 4 hours * £25 wage = £100 per observation, or £50 per participating teacher.

Initial Training Costs—The cost per teacher for initial training was approximately £238. That total combines direct costs paid by the research project, and the cost of each teachers' own time devoted to the training. We estimate one day of teacher time total for training, £200 = £25 * 8 hours, including the actual training session, preparation and follow-up tasks, and travel time. The total direct costs paid by the research project were £600 per treatment school or £37.5 per teacher. These costs include labor expenses for the research team members who developed and delivered the training, reimbursement for travel expenses of the trainers and some trainees, printed materials, and venue and catering expenses.

Software and Tablet Computers—Observers used tablet computers to record observation scores. Each school received 12 Apple iPads costing roughly £2,000 per school. The software ("app") was provided by RANDA Solutions at a cost of £210K total or £4,900 per school. Together £6,900 per school, or about £430 per teacher.

Cost per Teacher per Year—All together we estimate that the total cost of the peer evaluation program was just under £900 per teacher for the experiment's two-year intervention. The cost per teacher per year was thus just under £450, or about 1 percent of a £40K annual salary. This £450 includes the ongoing costs of the observations themselves. The total cost of observations is 2.27 observations * £50 = £113.50 per year or £227 total. And the £450 includes the annualized costs of initial training, £238, and computers and software, £430.

While £900 per teacher was the cost of the experiment's intervention, £900 is likely an overestimate of the cost per teacher as the program scales up or continues over time. The experiment was limited to math and English teachers, but the school's fixed costs—e.g., tablet computers and software licenses—could presumably be shared across other subjects. If schools continued peer observation for several years, the initial fixed costs—e.g., training for teachers and tablet computers—would fall as a share of annualized costs. Moreover, while the tablets and software facilitated observations, they may not be necessary to achieve the treatment effects we observe.