Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Teachers’ use of class time and student achievement

Simon Burgess [a], Shenila Rawal [b], Eric S. Taylor [c,*]

[a] University of Bristol, School of Economics, Mary Paley Building, Priory Road, Bristol BS8 1TX, United Kingdom
[b] Oxford Partnership for Education Research and Analysis, 7 Queens Square, Lyndhurst Road, Ascot, Berkshire SL5 9FE, England
[c] Harvard University and NBER, Gutman Library 469, 6 Appian Way, Cambridge, MA 02138, United States

ARTICLE INFO

JEL No:
I21
J24

Keywords:
Education production
Teacher performance

ABSTRACT

We study teachers’ choices about how to allocate class time across different instructional activities, for example, lecturing, open discussion, or individual practice. Our data come from secondary schools in England, specifically classes preceding GCSE exams. Students score higher in math when their teacher devotes more class time to individual practice and assessment. In contrast, students score higher in English if there is more discussion and work with classmates. Class time allocation predicts test scores separate from the quality of the teacher’s instruction during the activities. These results suggest opportunities to improve student achievement without changes in teachers’ skills.

Teachers’ choices and skills affect their students’ lives. Students assigned to more-effective teachers learn faster and, as a result, go on to greater success in adulthood. Yet, while evidence continually shows differences in teachers’ contributions to their students’ outcomes, evidence about why those contributions differ remains scarce. In particular, we still know little about the role of instructional practices. Here “practices” is shorthand for the choices teachers make about how to teach, and the extent to which they successfully carry out those choices. Practices are constrained by teaching skills but are not synonymous with skills.[1]

In this paper we examine novel data on teachers’ practices, combined with the subsequent test scores of their students. We study teachers and students in public (state) secondary schools in England. Specifically, math and English classes leading up to the General Certificate of Secondary Education (GCSE) exams typically taken at age 16. The data on practices were collected during classroom observations conducted by other teachers working in the same school. We describe several empirical results, some new to the literature on the economics of teachers and teaching.

Our primary focus is teachers’ choices about how to allocate class time across different instructional activities. First, teachers do make different choices. Some teachers spend much of class time using traditional direct instruction, including lecturing and the use of textbooks, while other teachers devote more class time to students working with their classmates or individual practice. For our study, classroom observers recorded which of twelve instructional activities the teacher used and for what amount of class time. Observers simply recorded what activities were happening without judging the appropriateness or quality of the activity. The list of activities, shown in Table 1, includes things like “lecturing or dictation,” “one to one teaching,” and “open discussion among students and teacher.” As an example, over one-third of teachers in our data used the activity “open discussion…” for most or all of class time, but one-quarter of teachers did not use any “open discussion…” Yet, while individual teachers choose different activities, those choices are largely unrelated to the subject being taught (English or math) or to the characteristics of the students in the class.

Second, teachers’ choices are (potentially) consequential for their students’ achievement. Students score higher on math exams when their teacher devotes more class time to individual practice and assessment. In the typical (average) math class, the teacher allocates “some of the time” to assessment and practice. In a class where the teacher allocates “most of the time” to practice and assessment, GCSE scores are 0.08σ

* corresponding author.
E-mail address: eric_taylor@gse.harvard.edu (E.S. Taylor).

[1] Jackson, Rockoff, and Staiger (2014) review the literature on teachers. Teachers and schools are not unique in this respect. As Syverson (2011) reviews, evidence from many sectors and industries shows large differences in productivity between firms, plants, etc., but the causes of those differences are only partially understood. Some intuitive potential causes—like “management practices”—get less attention in the literature because they are difficult to measure and difficult to test (quasi-)experimentally (on management see Bloom and van Reenen 2007, Bloom et al. 2013, and Bloom et al. 2015 for schools). Teaching practices are similarly difficult to measure, and difficult to manipulate (quasi-)experimentally.

https://doi.org/10.1016/j.econedurev.2023.102405
Received 2 November 2022; Received in revised form 28 March 2023; Accepted 11 April 2023
Available online 17 April 2023

Table 1
Instructional activities.

Activities used by observers	Activity group
1. Open discussion among students and teacher	Student peer interaction
2. Students are working in groups	
3. One to one teaching	Personalized instruction
4. Spending special time to assist weak students	
5. Students are doing written work alone	Practice and assessment
6. Gauging student understanding (e.g., through written or oral assessment)	
7. Assigning homework or class work to students	
8. Lecturing or dictation (one way transaction, teacher was speaking and students were listening)	Direct instruction
9. Students copying from the whiteboard	
10. Use of white board by teacher	
11. Teacher was using a textbook during teaching activities (use of examples from text, taking reference of text, read the lines of chapter)	
12. Engaged in non-teaching work (maintenance of register, preparation of data, format preparation etc.)	

Note: Activities list adapted from the SchoolTELLS project (Kingdon et al., 2008). The grouping of activities in the right hand column is described in Section 2.1.

higher than the typical class (σ = student test score standard deviations). For English exams, by contrast, students score higher when they spend more time working and talking with classmates, with the predicted gains similar in magnitude to math.

We need to interpret this result carefully. For comparison, we note that the path-breaking work on management practices and firm performance led by Bloom and van Reenen (2007) including work on school management (Bloom, Eifert, Mahajan, McKenzie & Roberts, 2013). Those papers have the same structure as this paper: collect data on agents (managers or teachers) about their actions at work (management tasks or classroom time use) and relate those actions to outcomes (firm performance or pupil achievement). The issue for both those papers and this paper is the extent to which we can control for confounders affecting the outcome measure. Student-level omitted variables are unlikely to bias our results. As is common in the literature on teacher performance, we address student-to-teacher selection by controlling for students' prior achievement or by using only within-student between-subject variation (Jackson, Rockoff & Staiger, 2014). Instead, the main concern is teacher-level omitted variables. For example, perhaps students learn more because their teachers are more skilled, and more skilled teachers choose different instructional activities. This kind of teacher-level omitted variables concern limits causal claims in many studies of teachers, even when students are randomly assigned to teachers (e.g., Aucejo, Coate, Fruehwirth, Kelly & Mozenter, 2020; Kane, Taylor, Tyler

Table 2
Rubric standards and associated description of "Effective."

Domain 1. Classroom Environment	
1.a 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.
1.b Establishing a Culture for Learning	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.
1.c Managing Classroom Procedures	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.
1.d Managing Student Behaviour	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.
1.e 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. Instruction	
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.
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 rigor. The lesson's structure is coherent, with appropriate pace.
2d Use of Assessment	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.
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 the *Framework for Teaching* (Danielson, 2007).

& Wooten, 2011; Taylor, 2018).

In fact, our data allow us to go a long way towards eliminating this problem because we have a measure of teacher instructional effectiveness. Observers rated the quality of teaching they observed using a detailed rubric, the *Framework for Teaching* (Danielson, 2007). These instructional effectiveness ratings measure a combination of skills and effort in ten teaching tasks, judged against a normative standard defined by the rubric. For example, one of the ten tasks is "using questioning and discussion techniques" (see Table 2).

Controlling for the teacher's instructional effectiveness gives us our third result: the instructional activities a teacher chooses predict her students' achievement independent of her teaching skills. We find the same patterns: time for individual practice and assessment benefits math scores, and peer interaction benefits English scores. The point estimates are one-quarter to one-third smaller but remain educationally meaningful and statistically significant at conventional levels. These results suggest that, separate from the teacher's skills or effort, some approaches to classroom instruction are more successful in promoting student learning than others. This result is perhaps this paper's most novel contribution. The identifying assumption for a causal interpretation is that our measure of teacher instructional effectiveness captures the teacher's skills which are correlated with her instructional activity choices and student achievement. Even if some degree of omitted variable bias remains, our estimates have a much stronger causal claim than existing estimates which entirely omit teachers' skills.

Fourth, a teacher's instructional effectiveness ratings also predict higher achievement scores. A student assigned to a top-quartile teacher, as measured by effectiveness ratings, will score about 0.08σ higher than a similar student assigned to a bottom-quartile teacher. That difference is roughly the same magnitude as the difference predicted by teachers' use of class time for practice in math or for peer interaction in English.

These relationships between teachers' practices and student test scores are educationally and economically meaningful. An improvement of 0.08σ is about one-third of the standard deviation in teachers' total contributions to GCSE scores (Slater, Davies & Burgess, 2012). Improvements in GCSE scores also predict future earnings and college going (Hayward, Hunt & Lord, 2014; Hodge, Little, & Weldon, 2021; Mcintosh, 2006).

This paper makes two contributions to the literature. Our primary contribution is demonstrating the relationship between a teacher's choice of instructional activities and her students' achievement, even conditional on her instructional effectiveness. Many papers measure differences between teachers in *how effectively* they do their work; classroom observation rubric ratings are quite common (see Jackson et al., 2014 for a review). Few papers have measures which can distinguish between what teachers *do* at work from *how effectively* they do it. The closest examples to this paper, of which we are aware, are Aslam and Kingdon (2011) and Taylor (2018). Both papers distinguish between instructional activities and teachers' skills, and both use those measures to predict student test scores, but the measures and settings are quite different from this paper.

This first contribution has important implications for teachers, and managers of teachers, working to improve schooling. Often the focus in schools, and among researchers, is on improving teachers' skills. Often those skills are difficult to learn, like managing student misbehavior or asking effective questions in class. The results in this paper suggest, for example, that students would learn more math if teachers simply spent more class time on individual practice, even without a change in teachers' skills.

A second contribution is new estimates of the correlation between observation rubric ratings—instructional effectiveness ratings—and student test scores. Several prior studies report the same correlation, and our estimates are similar in magnitude (for example, Kane & Staiger, 2012; Kane et al., 2011; Kane, McCaffrey, Miller & Staiger, 2013). Still, our estimates are novel in a few ways. First, they are from secondary schools in England. Existing estimates are almost entirely from elementary and middle schools in the United States; one exception is Araujo, Carneiro, Cruz-Aguayo and Schady (2016) who study kindergarten classes in Ecuador. Second, the observers in our study had little training compared to prior studies. Observers typically receive much more training, often including tests to insure inter-rater reliability. Third, rubric ratings may depend on the instructional activities used during the observer's visit. We can control for instructional activities when estimating the correlation with student test scores.

In the next section we describe the teachers, students, and schools in our study. Section 2 focuses on our measures of instructional activities and instructional effectiveness, and the observed differences in teachers' choices and skills. In Section 3 we examine the relationship between teachers' practices and their students' achievement test scores. We conclude in Section 4.

1. Setting and sample

We study teachers who work in public (state) secondary schools in England, and who teach math and English to year 10 and 11 students (roughly ages 14–16). Our measure of student achievement is GCSE exams, which students take at the end of year 11. Our measures of teaching practices—both instructional activities and instructional effectiveness—come from classroom observations conducted by coworker teachers.

The classroom observation data were gathered as part of a prior field experiment in the 2014–15 and 2015–16 school years. Full details and results of the experiment are described in Burgess, Rawal and Taylor (2021). The treatment schools began a new program of teacher peer observation. At each of the treatment schools, some teachers were always the observers, some always the observees, and some participated in both ways. Schools were randomly assigned to treatment or control, and teachers were randomly assigned to observer and observee roles. Section 2 describes the data collected in the peer observations. While teachers scored each other, the program did not involve any (formal) incentives or consequences linked to those scores.

All student data come from the UK government's National Pupil Database (NPD), including individual students' scores from the General Certificate of Secondary Education (GCSE) exams. At the end of year 11, students take GCSE exams in several subjects, but we use only math and English scores in this paper. The GCSE exams are high stakes for students; for example, scores influence college admissions. And GCSEs predict future earnings (Hayward et al., 2014; Hodge, Little, & Weldon, 2021; Mcintosh, 2006). Besides GCSE scores, the NPD data provide students' prior exam scores, demographics, and measures of exposure to poverty in their families and neighborhoods.

The NPD does not collect data linking students to their specific teachers. During the peer-observation experiment, schools provided class rosters which we use to link students and teachers. The rosters use masked teacher ID codes which, unfortunately, we cannot link to any other data on individual teachers.

Our study sample includes 251 teachers in 32 schools, and just over 7000 students who were taught by those teachers and for whom we have GCSE test scores. For math we have 5211 students and 136 teachers, and for English 4301 and 120.[2] The classroom observation data were collected by 231 different peer teachers.

Selection into this sample involved three steps. First, schools volunteered to participate in the new peer observation program experiment. The research team contacted nearly all high-poverty public (state)

[2] This subject difference is because math teachers were slightly more likely to be observed, not because we have differentially missing exam scores for students.

Table 3
Descriptive characteristics.

	Experiment schools	Schools with any observed teacher	Observed Teachers
	(1)	(2)	(3)
Prior English score	0.006	0.009	0.039
	(1.00)	(1.00)	(0.98)
Prior math score	0.007	0.008	0.058
	(1.00)	(1.00)	(0.97)
Female	0.487	0.488	0.480
IDACI	0.276	0.279	0.314
	(0.17)	(0.17)	(0.18)
Ever free school meals	0.398	0.402	0.426
Birth month (1–12)	6.569	6.579	6.581
	(3.42)	(3.42)	(3.39)
London school	0.162	0.164	0.180

Note: Means and standard deviations (in parentheses) for the samples described by the column headers.

secondary schools in England and invited them to participate in the experiment.[3] Schools were not selected based on student test scores. Second, half of volunteer schools were randomly assigned to the treatment program. Third, within each of the treatment schools, a random sample of teachers were selected to be observed and scored. Fourth, teachers chose whether or not to participate. Thus, our sample of 32 schools and 251 teachers is partly randomly selected and partly self-selected.[4]

Table 3 provides a description of our sample. Schools invited to participate in the experiment were intentionally selected to have high poverty rates, and that initial selection is reflected in the Income Deprivation Affecting Children Index (IDACI) and free school meals rows of Table 3. Just over 40 percent of students are, or ever have been, eligible for free school meals, substantially higher than the national average. Comparing across the columns of Table 3 provides some information on teacher self-selection into our sample.

2. Measures of teaching practices

Classroom observations revealed meaningful differences between teachers in both the instructional activities teachers chose to use in class, and in rubric-based ratings of teachers' instructional effectiveness. In this section we describe the variation in practices and effectiveness. In the next section we relate teacher measures to student test scores.

The observation data were collected during nearly 2700 classroom visits, where one observer scored one of her peer teachers. Visits typically lasted 15–20 min. The typical (median) teacher was observed eight times over two years (IQR 4–15). The typical teacher was scored by three different peer observers (IQR 2–5). All teachers received training on the rubric and other aspects of the program. However, the training was brief in comparison to the training observers have received in other studies and settings (e.g., Kane & Staiger, 2012; Kane et al., 2011).

2.1. Instructional activities

To measure teachers' instructional activity choices, observers were given a list of activities and asked to record how frequently each activity was used. Importantly, peer observers recorded only the frequency of the activity during their visit; observers were not asked to assess the quality or appropriateness of the activity. The complete list of twelve activities is shown in Table 1, including things like "open discussion among children and teacher" and "use of white board by teacher." Observers could choose from five options: none (0), very little (1), some of the time (2), most of the time (3), full time (4). The activities list and instrument were adapted from the SchoolTELLS project (Kingdon, Banerji & Chaudhary, 2008).

Teachers make quite different choices about how to spend class time. Fig. 1 shows the twelve different instructional activities and the frequency of their use. For example, in more than one-third of classes observers recorded "open discussion among children and teacher" during most or all of the class time. Yet, in one-quarter of classes "open discussion…" was very rare or entirely absent. Teachers were similarly split on "children doing written work alone." A contrasting example is use of a textbook, which was recorded as rare or absent in nearly nine out of ten classes.

The patterns of instructional activities are quite similar in math and English classes. The correlation between subjects in the average frequency of activities is 0.96. Appendix Fig. A1 shows Fig. 1 separately by subject. However, this similarity of time use does not mean the activities contribute to students' math and English test scores in the same way, as we show in Section 3.

Some groups of instructional activities are correlated. Table 4 shows the correlation matrix for the twelve activities. Some activities may be complementary inputs to student learning, while other activities can occur simultaneously for practical reasons. Examining the correlations, together with the substance of the activities, suggests an opportunity for dimension reduction.

Our analysis focuses on four groups of instructional activities. First, a group we label "student peer interaction," combining activities 1–2, which involve students interacting with each other (and the teacher). Second, "personalized instruction," combining activities 3–4, which involve personalized attention from the teacher to students. Third, "practice and assessment," combining activities 5–7, which involve student practice and assessment. Fourth, "direct instruction," combining activities 8–11, which involve traditional lecturing and other direct instruction. To measure each activity group, we use the simple average of the items within the group.

Table 5 describes teachers choices using these four groups of activities. The most common activity group is "student peer interaction" with a mean of 1.7, where a 2 is "some of the time" on the scale of 0 "not at all" to 4 "full time." The least common is "direct instruction" with a mean of 1.2. Most of the variation in these activities is between teachers within schools; differences between schools account for 15–30 percent of the variation in activity frequency. Teachers do combine the four activity types in class, with correlations between 0.20–0.40 in our observation data. Appendix Fig. A2 provides additional detail on how teachers combine the activities: We show the frequency of each activity among math teachers who use "practice and assessment" the most versus

[3] For this purpose "high-poverty schools" were those schools where the percent of students eligible for free school meals was above the median for England.

[4] There is another dimension of sample selection: Our observation data are (potentially) a non-random sample of teachers' behavior. The non-randomness could arise, for example, through the timing of visits or through teachers' distorting their behavior while being evaluated. We return to these topics in Section 3 as potential threats to our interpretation of the results.

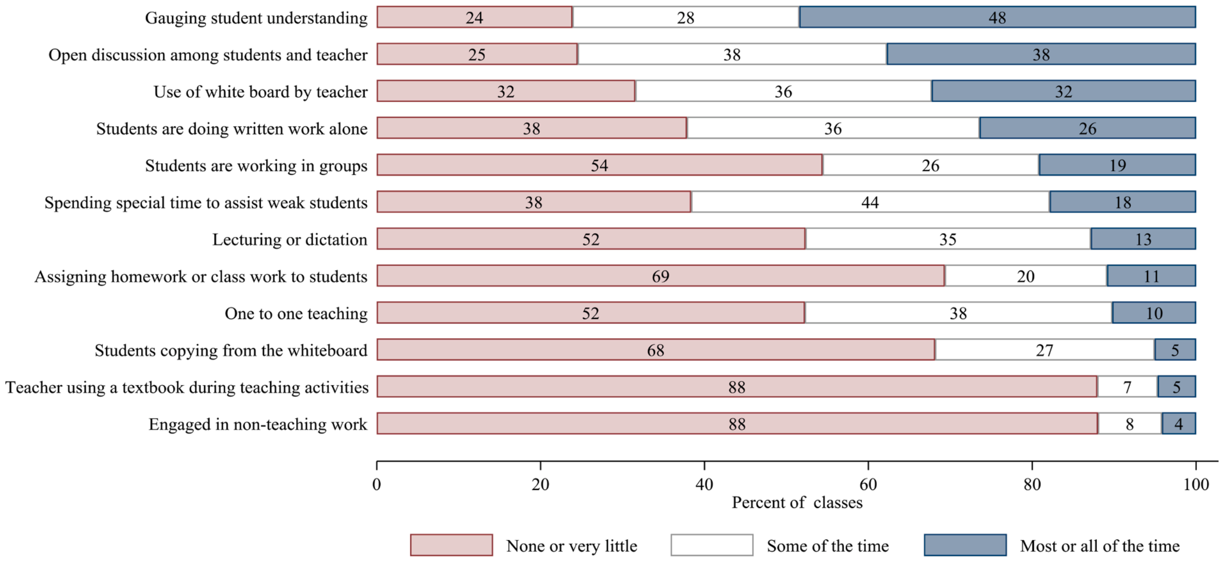

Fig. 1. Frequency of instructional activities.
Note: For each activity, the red (left) bar is the proportion of classes where there was "none" or "very little" of the activity. The blue (right) bar is the proportion of classes where the activity was occurring "most of the time" or "full time." The white (middle) bar is the "some of the time." Proportions are of 2687 observations, each the visit of a peer observer k to the class of teacher j.

Table 4
Correlations among instructional activities.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Open discussion among students and teacher	1										
2. Students are working in groups	0.19	1									
3. One to one teaching	0.01	0.11	1								
4. Spending special time to assist weak students	0.08	0.14	0.39	1							
5. Students are doing written work alone	−0.09	−0.12	0.17	0.14	1						
6. Gauging student understanding	0.24	0.17	0.09	0.17	0.22	1					
7. Assigning homework or class work to students	0.08	0.14	0.09	0.14	0.20	0.23	1				
8. Lecturing or dictation	−0.09	−0.11	−0.04	−0.08	0.04	−0.02	0.12	1			
9. Students copying from the whiteboard	0.00	−0.06	−0.01	0.01	0.07	0.03	0.12	0.31	1		
10. Use of white board by teacher	0.05	−0.08	−0.03	0.00	0.00	0.13	0.09	0.29	0.33	1	
11. Using a textbook during teaching activities	−0.04	0.04	0.07	0.05	0.10	0.04	0.15	0.10	0.19	0.04	1
12. Engaged in non-teaching work	0.00	0.07	0.08	0.05	0.20	0.08	0.26	0.07	0.13	0.05	0.20

Note: Correlations of class time use among twelve instructional activities, net of observer effects. Each of the 2687 observations is the visit of a peer observer k to the class of teacher j. Observers recorded time use in five ordered categories: (0) none, (1) very little, (2) some of the time, (3) most of the time, and (4) full time. Before estimating the correlations, we first calculate observer k's mean for each item and subtract that mean from all scores k assigned for that item.

the least (top versus bottom quartile), and for all other combinations of activities and quartiles. In Section 3 we show that these four activity types predict student scores quite differently in math compared to English.

Simplification involves tradeoffs. Our grouping divides the twelve activities into mutually exclusive and exhaustive categories which are relatively straightforward. The tradeoff is that these simple groups ignore variation in how activities are correlated within and between groups. To complement the simple grouping, we show in online Appendix C that the paper's results are robust to using principal components analysis for dimension reduction.

Finally, the activity data were collected using a scale which is not a strongly interval scale: none (0), very little (1), some of the time (2), most of the time (3), full time (4). Our goal in this paper is to understand how these activity inputs predict student achievement score outcomes. In such cases non-interval predictor variables increase the risk of mistaken conclusions about nonlinear relationships and about extrapolations far away from the support of the data. We limit our estimates to the best linear prediction and limit our interpretation to changes near the mean of each activity measure. However, there is some empirical evidence which supports treating our activity data as interval scaled. Table 4 reports polychoric correlation estimates, which relax the assumption of interval scaled data, but these are very similar to the conventional Pearson correlation estimates.

2.2. *Instructional effectiveness*

To measure a teacher's instructional effectiveness, observers rated teachers using a structured rubric known as the *Framework for Teaching* (Danielson, 2007, "FFT"). The rubric is widely used by school systems and in academic research. Teachers are rated on ten separate instructional tasks (or "standards" in the FFT jargon), which are listed in the left-hand column of Table 2. For each task, the rubric includes detailed descriptions of what observed behaviors should be scored as "highly effective" teaching, "effective," "basic," and "ineffective." In Table 2 we reproduce the descriptions for an "effective" rating, as an example. The

Table 5
Instructional activities.

	Correlation matrix Pooled				Mean (st.dev.)		
					Pooled	Math	English
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Direct instruction	1				1.23 (0.71)	1.36 (0.72)	1.06 (0.66)
Student peer interaction	0.17	1			1.68 (0.93)	1.71 (0.91)	1.66 (0.94)
Personalized instruction	0.21	0.31	1		1.44 (0.92)	1.52 (0.91)	1.34 (0.92)
Practice and assessment	0.37	0.28	0.35	1	1.58 (0.91)	1.73 (0.86)	1.38 (0.95)

Note: Means and standard deviations (columns 5–7) for, and correlations among (columns 1–4), class time use in four groups of instructional activities, described by row labels. This table uses a sample of 2687 observations. Each of the 2687 observations is the visit of a peer observer *k* to the class of teacher *j*. Each of the four measures (rows) is itself the average of several item level scores recorded by peer observers, as described in the text. Time use is measured in ordered categories: (0) none, (1) very little, (2) some of the time, (3) most of the time, and (4) full time.

Table 6
Instructional effectiveness ratings.

	Pooled (1)	Math (2)	English (3)
Overall average	9.09 (1.75)	9.15 (1.80)	9.00 (1.69)
Classroom environment average	9.27 (1.84)	9.35 (1.88)	9.17 (1.78)
1a. Creating an environment of respect and rapport	9.32 (2.04)	9.35 (2.09)	9.28 (1.97)
1b. Establishing a culture for learning	9.20 (2.01)	9.25 (2.04)	9.13 (1.96)
1c. Managing classroom procedures	9.24 (2.04)	9.31 (2.06)	9.14 (2.01)
1d Managing student behavior	9.41 (2.05)	9.42 (2.12)	9.41 (1.96)
1e. organizing physical space	9.13 (2.18)	9.29 (2.14)	8.87 (2.23)
Instruction average	8.90 (1.83)	8.94 (1.87)	8.86 (1.77)
2a. Communicating with students	9.29 (1.91)	9.31 (1.95)	9.25 (1.85)
2b. Using questioning and discussion techniques	8.77 (2.17)	8.80 (2.16)	8.72 (2.18)
2c. Engaging students in learning	8.99 (2.00)	9.03 (2.09)	8.93 (1.86)
2d Use of assessment	8.50 (2.21)	8.53 (2.19)	8.46 (2.23)
2e. Demonstrating flexibility and responsiveness	8.83 (2.05)	8.78 (2.08)	8.90 (2.01)

Note: Means and standard deviations (in parentheses), using a sample of 2687 observations in column 1. Each of the 2687 observations is the visit of a peer observer *k* to the class of teacher *j*. The samples for columns 2 and 3 are 1510 and 1177 respectively. For each of the ten numbered items above, observers rated effectiveness on a 1–12 scale: 1–3 ineffective, 4–6 basic, 7–9 effective, and 10–12 highly effective. The three average scores above are the mean of the relevant item level scores, ignoring missing scores.

full rubric is provided in online Appendix B.[5]

Teachers do differ in instructional effectiveness, as rated by their coworkers. To be clear, in this context "instructional effectiveness" is a measure of a teacher's observable actions in the classroom. While we use the word "effectiveness," these ratings could also be described as measuring "job performance." The ratings reflect a combination of a teacher's skills and effort applied to specific teaching tasks, judged against a normative standard defined by the rubric.

For each of the ten instructional tasks, Table 6 reports the mean rating and standard deviation. In this study peer observers assigned a score from 1 to 12 to each of the ten rubric items. In most settings the FFT rubric is scored 1–4 corresponding to the four descriptions. Our observers were trained to use scores of 1–3 for "ineffective," 4–6 for "basic," 7–9 for "effective," and 10–12 for "highly effective." Thus, for example, an observer who felt the teacher was "effective" could chose a score of 7, 8, or 9, with 7 suggesting "effective" but closer to "basic" and 9 suggesting "effective" but closer to "highly effective."

Observers rated teachers highest, on average, for "managing student behavior" (mean 9.4) and lowest for "use of assessment" (mean 8.5). In general, teachers were rated more effective in classroom environment

[5] These ten scored tasks (or standards) are divided into two "domains" of "classroom environment" and "instruction." These two domains are scored during in-class observations, and during the peer-evaluation experiment only these two domains were scored. The FFT rubric also includes several standards in two other domains, "planning" and "assessment," which are scored based on conversations with the teacher and a review of materials.

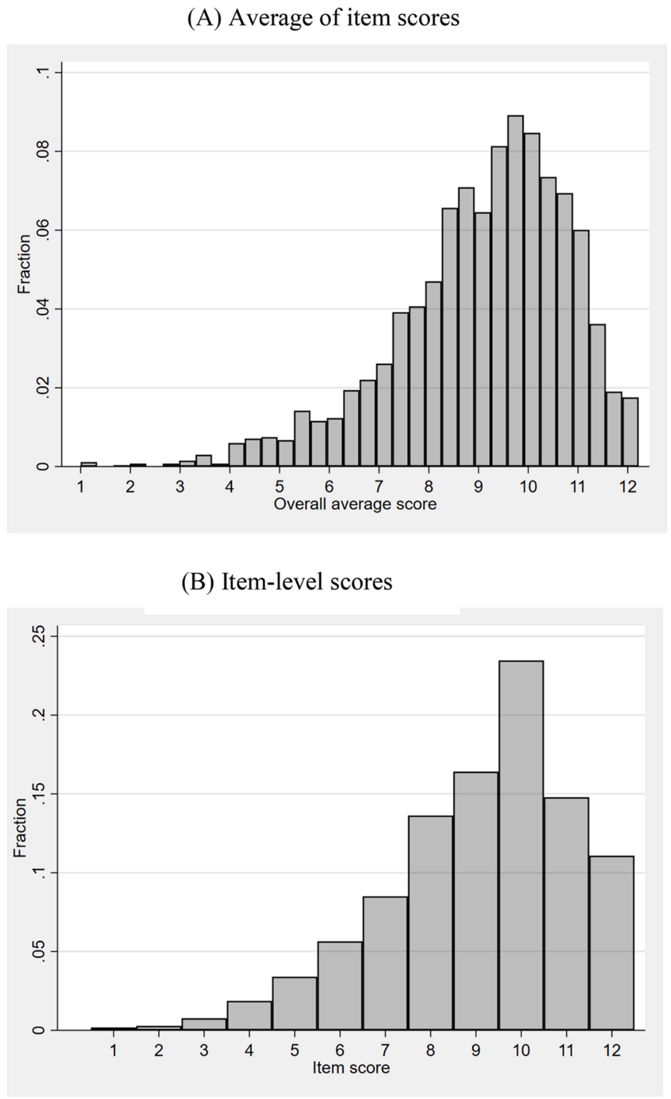

Fig. 2. Distribution of instructional effectiveness ratings.
Note: Panel A shows a histogram of 2687 average scores. Each of the 2687 observations is the visit of a peer observer *k* to the class of teacher *j*. The x-axis is the simple average of the ten item scores for a given observation visit, ignoring missing item scores. Panel B shows a histogram of the 23047 item-level scores recorded across the rubric's ten items.

tasks than instruction tasks. "Use of assessment" also showed the largest differences in effectiveness between-teachers (standard deviation 2.2). Teachers were most similar in "communicating with students" (standard deviation 1.9).

A teacher's instructional effectiveness ratings across the ten tasks are strongly correlated. The average correlation in rating between any two tasks is 0.70, with a range of 0.55 to 0.86. The full matrix of pairwise correlations is shown in Appendix Table A1. The correlation across tasks partly reflects the fact that the true underlying skills and efforts are correlated. The correlation also partly arises because the ten task ratings are given by one observer. However, using only within observer variation the average pairwise correlation is still 0.60, with a range of 0.44 to 0.79.

In practice, then, the rubric ratings mostly measure one general dimension of instructional effectiveness. Appendix Table A2 shows results from a principal components analysis of the item-level ratings. The first principal component is effectively the simple average of the ten task items, and that simple average explains three-quarters of the variation in the item-level ratings. For comparison, the first principal component of instructional activities items explains only 13 percent of the activities data. This pattern of correlations among ratings matches prior studies using the FFT observation rubric.[6]

Given these correlations, we focus on a single score for instructional effectiveness: the simple average of the ten FFT rubric ratings. The top panel of Fig. 2 shows a histogram of these average scores, with one score for each observation.

The scores in our data may not fully reflect the true differences between teachers in their instructional effectiveness. Here we describe three sources of potential measurement error. Later in Section 3 we discuss how these sources of error might affect our interpretation of the paper's main results. First, the scores are based on only a sample a teacher's instruction. Recall that the average teacher was observed four times a year, with each observation lasting 15–20 min. Thus, we would expect some classical sampling error. In our data, 36 percent of variance of the observation-specific FFT scores is persistent differences between teachers; the reliability of scores based on one observation per teacher is 0.36 and based on four observations 0.68. These estimates are quite similar to what Kane and Staiger (2012) and Ho and Kane (2013) found for FFT scores in the Measures of Effective Teaching (MET) Project.

Second, classroom observation ratings often have a skewed distribution with ceiling effects. This pattern can be seen in Fig. 2 for our data. One explanation is that the rating scale may be less-sensitive to true performance differences at the top of the distribution. However, often this pattern of teacher ratings is interpreted as leniency bias (Kraft & Gilmour, 2017; Weisberg et al., 2009), and leniency bias is a common feature of performance evaluations in many occupations (Prendergast, 1999). We might predict greater leniency bias in our setting because observers and observees worked together in the same school as peers. Alternatively, we might predict less leniency bias because of the low-stakes nature of the peer observations.

These common patterns—skew and ceiling effects—are much weaker in our data. The bottom panel of Fig. 2 shows a histogram of item-level ratings, pooling together all ten tasks. The most common score is 10 out of 12, given in almost 25 percent of ratings; but scores of 8, 9, 11, and 12 each have 10–15 percent of ratings. Still, as in many other settings, very few teachers are scored 3 or below ("ineffective"). Moreover, these common patterns are further weakened by using the average score, as shown in the top panel of Fig. 2. In the end, while ratings are bunched at the top of the scale in our data, there is more variation than is typical of classroom observations and much less of a ceiling effect than is typical.[7]

Third, observers received training on the FFT rubric, but their training was brief compared to training in other studies. The lighter training may have reduced inter-rater reliability, though overall reliability of our FFT scores was not different from prior studies. Additionally, without the goal of inter-rater reliability, observers may have been more likely to be influenced in their ratings by information learned outside of the formal observation visit. Ho and Kane (2013) report evidence suggesting some school principals use outside information when rating teachers.

Finally, a growing evidence base supports the validity of using FFT scores to make inferences about a teacher's skills and her effects on

[6] In three prior studies in U.S. schools, ratings for the ten FFT items are correlated 0.72-0.88 (Kane & Staiger, 2012; Ho & Kane, 2013; Gitomer et al., 2014; ICPSR n.d., Andrew Ho personal communication May 3, 2019). Kane et al. (2011) reports similar principal components results. However, in our data rating levels are consistently higher across items, about 0.9 points on the 4-point scale, and our ratings have higher variance, about 30 percent larger.

[7] The variation is likely due in part to using a 12 point scale, instead of the conventional 4 point scale. Appendix Fig. A3 shows a histogram of the same data as Fig. 2 panel B, but where the 1-12 ratings are collapsed into the more-common 1-4 scale. The skew and ceiling effects are, not surprisingly, much stronger.

student achievement. First, a teacher's FFT scores predict her scores on alternative measures of teaching skills. In the MET Study, FFT scores were correlated roughly 0.70–0.90 with scores from four other observation rubrics (Kane & Staiger, 2012). Second, a teacher's FFT scores predict her value-added contribution to her student's achievement test scores. In the MET Study, FFT scores were correlated 0.13–0.19 with value-added scores. While that correlation may be relatively small, the implied effect is educationally meaningful. A student in the classroom of a top-quartile FFT teacher would gain the equivalent of an extra 1.5 months of math instruction, compared to being taught by the average teacher (Kane & Staiger, 2012; Kane et al., 2013). We find a similar correlation in this paper (see Section 3.4), as do Kane et al. (2011). Third, prior (quasi-)experiments show that exposing teachers to the FFT as a treatment improves value-added (Taylor & Tyler, 2012; Burgess et al., 2021).[8]

2.3. Teaching practices and student types

One last note on measuring teaching practices. Observed differences between teachers may partly reflect differences in the students they teach. Teachers may choose different instructional activities for students with different academic needs, or students with different needs may be assigned to teachers based on the teacher's instructional effectiveness or use of activities. Such intentional choices or assignments may or may not improve a school's success (Duflo, Dupas & Kremer, 2011; Ballatore & Sestito, 2016; Aucejo et al., 2020; Graham, Ridder, Thiemann & Zamarro, 2021). Alternatively, the judgements of classroom observers may be influenced by the students in the class during the visit (Campbell & Rondfeldt, 2018).

However, in this paper's setting, we find little evidence of a relationship between students' observable characteristics and their teacher's instructional practices. Appendix Table A3 reports estimates from regressions where the outcome is a student or class characteristic—prior test score, exposure to poverty, class average prior score, etc.—and the predictors are our measures of time use across different class activities. We find no meaningful pattern of correlation between students and activities. The same conclusion is true when we predict student characteristics using our FFT measure of instructional effectiveness. In short, teachers' instructional choices do not appear to depend on the students they are assigned.

3. Teaching practices and student achievement

We now turn to the relationship between teaching practices and student achievement. As we detail in this section, students score higher in math when their teacher allocates more class time to student practice and assessment. By contrast, students score higher in English when teachers give more time to students working and talking with each other in class. These relationships—between instructional activities in class and student achievement—hold even controlling for the teacher's instructional effectiveness. Students also score higher when their teacher is rated higher on instructional effectiveness.

3.1. Estimation

Our estimation strategy begins with a conventional statistical model of student test scores:

$$A_{ijs} = T_j\delta + X_{ijs}\beta + \lambda_s + \varepsilon_{ijs} \tag{1}$$

where A_{ijs} is the standardized GCSE score for student i in subject s (math or English) taught by teacher j in the school year leading up to the GCSEs.[9] The vector T_j represents scores or measures taken from the classroom observations of teacher j, and described in Section 2. Our interest is in estimating δ. The vector X_{ijs} includes several additional controls: student i's own prior test scores in math and English; the class means and standard deviations of the two prior test scores, leaving out i; and several other student observables.[10] The λ_s term represents subject fixed effects.

Our preferred estimates of δ also account for differences between observers. Building on specification 1, we fit:

$$A_{ijks} = T_{jk}\delta + X_{ijs}\beta + \lambda_s + \theta_k + \epsilon_{ijks} \tag{2}$$

where T_{jk} is the scores given to teacher j by observer k. The addition of observer fixed effects, θ_k, controls for differences between observers in their expectations, practices, experience, etc.[11] To estimate specification 2, we first create a new data set with K_j duplicates of each student-teacher pair record, ijs, in the original data, where K_j is the number of observers who scored teacher j. To these new data we add the T_{jk} scores.[12] We then estimate 2 weighting by $1/K_j$; thus, each student-teacher pair, ijs, is given equal weight regardless of number of observers who rated teacher j. Throughout the paper we report cluster robust standard error estimates, where the clusters are teachers j.[13]

We report estimates of δ separately by subject. We estimate specification 2 but allow all δ and β terms to be different by subject. Observer fixed effects, θ_k, remain cross subject for our main results, but those results are robust to using observer-by-subject fixed effects (equivalently, estimating 2 separately by subject).

3.2. Instructional activities in class and student test scores

Different teachers choose to allocate class time in different ways—lecture, group discussion, individual practice, etc.—and those different instructional activities partly explain differences in student test scores. As the estimates in Table 7 column 1 panel A show, students score higher on the math GCSEs when their teacher's approach includes more time for individual practice and assessment. Increasing time for "practice and assessment" by one standard deviation predicts 0.068σ higher math test scores. In the typical (average) math class, the teacher allocates "some of the time" to assessment and practice. In a class where the teacher allocates "most of the time" to practice and assessment, we would expect scores to be roughly 0.08σ higher than the typical class.[14]

[8] These predictive validity results are complemented by the history of the FFT. The research used to design the FFT began in the 1990s at the Educational Testing Service for the PRAXIS III, including developing a detailed theoretical framework for how teaching affects learning (Dwyer & Villegas, 1993; Myford et al., 1994; Danielson, 2007).

[9] Strictly speaking the s index on A_{ijs} and ε_{ijs} is redundant because, in our data, every student is assigned to just one teacher per subject and thus $s = s(ij)$. We maintain the s index to facilitate the exposition. Student scores are standardized (mean 0, s.d. 1) by subject and school year within our analysis sample.

[10] Prior test scores are Key Stage 2 (KS2) scores. The other characteristics are gender, ever eligible for free school meals, IDACI score, birth month, and the year the student took the GCSEs. We also include an indicator for whether the school is in London.

[11] These observer differences might include, for example, differences between observers in their sense of what constitutes "gauging student understanding" or "non-teaching work," or the thresholds between "some of the time" and "most of the time."

[12] When k observes j more than once, we use the average measures or scores from k in T_{jk}.

[13] The primary motivation for the cluster (teacher) correction is that teachers' choices about class time use are the "treatment," for which we would like to know the effect on student learning. The correction is also motivated by the duplication of records required for the observer fixed effects.

[14] As shown in Table 5, the math mean for "practice and assessment" is 1.73 where 2 is "some of the time." The standard deviation is 0.86, thus a one-scale-point change from 2 "some of the time" to 3 "most of the time" would be roughly 0.068σ/0.86 = 0.08σ.

Table 7
Instructional activities and student achievement scores.

	(1)	(2)	(3)	(4)	(5)	(6)
(A) Math						
Instructional activities						
Direct instruction	0.012	0.006	0.023	0.009		
	(0.016)	(0.007)	(0.015)	(0.007)		
Student peer interaction	0.020	0.007	0.002	0.001		
	(0.013)	(0.006)	(0.014)	(0.008)		
Personalized instruction	0.004	0.003	0.003	0.003		
	(0.019)	(0.008)	(0.019)	(0.008)		
Practice and assessment	0.068**	0.023**	0.047*	0.015+		
	(0.019)	(0.008)	(0.019)	(0.008)		
Instructional effectiveness			0.070**	0.024*	0.077**	0.026**
			(0.018)	(0.010)	(0.017)	(0.010)
(B) English						
Instructional activities						
Direct instruction	−0.018	−0.024+	−0.009	−0.019		
	(0.020)	(0.013)	(0.019)	(0.012)		
Student peer interaction	0.053**	0.028**	0.043**	0.024**		
	(0.016)	(0.009)	(0.016)	(0.009)		
Personalized instruction	−0.021	−0.011	−0.026+	−0.014		
	(0.013)	(0.009)	(0.014)	(0.009)		
Practice and assessment	−0.024	−0.015	−0.030+	−0.021+		
	(0.016)	(0.011)	(0.017)	(0.012)		
Instructional effectiveness			0.039+	0.027*	0.040*	0.026*
			(0.021)	(0.011)	(0.019)	(0.010)
Student covariates	√		√		√	
Student fixed effects		√		√		√

Note: Point estimates and cluster (teacher j) corrected standard errors from several least-squares regressions, each with the same estimation sample of 253 teacher observations and 9512 student-by-subject observations. Each column reports estimates from a single regression. The dependent variable is a test score for student i in subject s (maths or English) measured in student standard deviation units. The key independent variables—the rows in the table—are observation scores for student i's teacher j in subject s, where $j = j(is)$. Teacher scores are measured in teacher standard deviation units. Teacher j's scores do not vary across students but do vary across the observers k who determined the scores. The data used to fit each regression are student i by teacher j (equivalently subject s) by observer k, but each ij pair is weighted equally, i.e., weighted $1/K_j$ where K_j is the number of observers k who scored teacher j. All specifications include observer k fixed effects, and indicator variables for subject. "Student covariates" include controls for student i's prior test scores in both subjects, gender, eligibility for free school meals, IDCACI score, month of birth, test year, and schools in London. All specifications include controls for the class mean and standard deviation of prior scores in both subjects. When a covariate is missing, we fill it in with zero, and include an indicator = 1 for missing on the given characteristic. + indicates $p<0.10$, * 0.05, and ** 0.01.

By contrast, the other class activities are much weaker predictors of math scores.

For English GCSEs, however, students score higher when class time includes more student interaction with their classmates. The coefficient on "student peer interaction" is 0.053σ for predicting English test scores (Table 7 column 1 panel B), roughly as large as "practice and assessment" is for math. But other activities are not strong predictors of English scores, indeed, more time in the other activities may lead to lower test scores.

The relationship between instructional activities and student achievement is economically and educationally meaningful. An improvement of 0.08σ is about one-third of the standard deviation in total teacher contributions to student test scores.[15] A difference of 0.08σ is also roughly the difference between being assigned to a first-year teacher or fifth-year teacher (see Jackson et al., 2014 for a recent review). A gain of 0.08σ is also similar to the gain from adding 2–3 weeks of instruction to the school year (Aucejo & Romano, 2016; Fitzpatrick, Grissmer & Hastedt, 2011; Sims, 2008).

The estimates in Table 7 column 1 alone are not sufficient to conclude that more practice and assessment causes higher math scores per se, or that more student interaction causes higher English scores. It is certainly plausible that students learn more or less because of how their teachers allocate class time. However, in column 1, we cannot rule out an alternative explanation: that students learn more or less because of something else their teachers do, and that something else is simply correlated with the teacher's choice of instructional activities. This teacher-level omitted variables concern also limits causal claims in other similar research, even when students are randomly assigned to teachers (e.g., Aucejo et al., 2020; Kane et al., 2011; Taylor, 2018).

One important potential omitted variable is the teacher's skill or effort. Whether or not a given instructional activity benefits student learning should depend, at least to some extent, on the teacher's skill or effort in that specific activity. Consider, for example, "student peer interactions" which includes open discussions among the class. Perhaps this activity contributes to higher achievement in English but not math, as in Table 7, because English teachers are more skilled in (or give more effort to) "using questioning and discussion techniques." Perhaps math teachers allocate more class time to "practice and assessment" because they are better at those tasks.

We test for this potential bias by adding instructional effectiveness ratings as controls, and examining whether and how the coefficients on instructional activities change. As discussed earlier, our measure of instructional effectiveness is a composite of teaching skills and effort.

Compare Table 7 columns 1 and 3. For math the point estimate for "practice and assessment" shrinks about one-third to 0.047σ, but is still meaningful, and we cannot reject the null that it is unchanged from 0.068σ. The point estimate for "direct instruction" nearly doubles to 0.023σ. This would be consistent with lecturing being more productive

[15] Slater, Davies, and Burgess (2012) estimate the standard deviation of teacher contributions to GCSE scores is 0.272 student standard deviations. This estimate comes from English secondary schools and GCSE courses, as in our current study, though the sample in Slater, Davies, and Burgess (2012) is broader. For a general summary of estimates on the teacher value-added distribution see Jackson, Rockoff, and Staiger (2014) and Hanushek and Rivkin (2010), though many of those estimates come from elementary and middle schools in the United States. The 0.272 estimate may be larger than other estimates in part because students typically spend two years with their GCSE teacher.

when the teacher is more skilled in math (see for example Taylor, 2018). For English, similar to math, the key point estimate on "student peer interaction" shrinks by about one-fifth but remains educationally meaningful. The negative coefficients on "personalized instruction" and "practice and assessment" become larger in absolute value and are somewhat more precisely estimated.

These results suggest that, separate from the teacher's skills or effort, some approaches to classroom instruction are more successful in promoting student learning than others. The results also suggest that the nature of effective activities may depend on the subject being taught. This result is perhaps this paper's most novel contribution to the literature. Research which combines both measures of instructional activities and measures of teacher skill to predict student test scores are rare. The closest, of which we are aware, are Aslam and Kingdon (2011) and Taylor (2018).

3.3. Additional considerations

Our estimation strategy also addresses potential student-level omitted variable bias, arising from how students are assigned to teachers. However, the threat of unobserved student characteristics is likely much less than the threat of unobserved teacher characteristics. Our estimates control for students' prior test scores, the distribution of peer prior scores, student backgrounds, and school effects. One limitation is that our prior test scores are Key Stage 2 tests taken five years prior to the GCSE tests, not the immediate prior school year. Well known evidence—from Chetty, Friedman and Rockoff (2014), Kane and Staiger (2008), and others—suggests it is plausible to assume student-teacher assignments are ignorable, in the causal inference sense, conditional on prior year test scores. It is less clear how the benefits for lagged score controls degrade with longer lags, though recent work in Angrist et al. (in-press) suggests reasons for optimism.

As an alternative approach, the even numbered columns in Table 7 use student fixed effects. The point estimates are smaller, shrinking by half to two-thirds. Still, the same pattern remains: math benefits from class time for individual practice and assessment, English benefits from student-peer interaction.

The lagged dependent variable estimates and student fixed effects estimates provide bounds on the influence of student-level omitted variable bias, like unobserved prior-year achievement. Correctly purged of any bias arising from the non-random sorting of students to teachers, the coefficient on "practice and assessment" for math would be between 0.015σ and 0.047σ, for example. The student fixed effects estimates will be correct—in the specific sense of avoiding any bias from omitted student characteristics—only if student-teacher assignments for both math and English are based on the same information. Otherwise, the student FE estimates are likely too small. Math and English assignments are made concurrently, and so the same information is available to the school for both decisions, but in practice the school may use different information in the two decisions.[16]

Even after accounting for the primary threats—teacher skills and student-teacher assignments—there may still be other omitted variables. One potential omitted variable is the timing of observation visits during the school year. For example, imagine that (i) all math teachers allocate more class time to "practice and assessment" later in the school year as the GCSE exam dates approach, but that (ii) teachers who make larger value-added contributions to student achievement scores are more likely to be observed later in the school year. If both (i) and (ii) are true, then we would find a positive correlation between "practice and assessment" and GCSE math scores in our data, even though all teachers use class time the same way. However, we find no evidence of this observation-timing threat in our setting. In Appendix Table A4 we show that the pattern of results in Table 7 does not change if we include month of observation effects. Additionally, Appendix Fig. A4 shows that class time allocation does not change over the school year.

A second example of a potential omitted variable is teacher conscientiousness. A more conscientious teacher may be more likely to distort her time use choices while being observed, choosing class activities she believes are the socially desirable activities among her peers. To bias our results, that same kind of conscientiousness would also need to increase the teacher's value-added contribution to test scores. Otherwise, the distorted data from a small sample of class time would bias against finding any relationship. Rockoff, Jacob, Kane and Staiger (2011) report no significant relationship between teacher value-added and conscientiousness, as measured with a Big Five instrument. Moreover, it is not obvious how a social-desirability motivated conscientiousness would improve value-added. Additionally, to create the pattern of results in Table 7, the socially-desirable class activities would have to differ by subject. Appendix Fig. A1 shows that, on average, teachers do not differentiate activities by subject when being watched by their peers.

Finally, as detailed in online Appendix C, our results are robust to changing how we group activities. In the alternative approach, we use principal components analysis to reduce the dimensionality of the instructional activity data. Then we repeat the analysis in Table 7, replacing the four activity groups with five principal component scores. The results show the same substantive patterns as Table 7; the substantive patterns are not an artifact of how we go about combining activity data. For example, in English, the fourth principal component is the stand-out predictor. This component, which we label "group vs. individual work," is increasing in activities where students interact with their classmates and decreasing in activities where students work alone or one-on-one with the teacher.

3.4. Teaching effectiveness ratings and student test scores

Rubric-based teaching effectiveness ratings also predict student GCSE test scores. In Table 7 column 5, the estimated coefficient on FFT score is 0.077σ for math and 0.040σ for English. Imagine two students: the first student is assigned to a top-quartile teacher, as measured by the FFT rubric, and the second to a bottom-quartile teacher. The first student will score more than 0.10σ higher than the second student on the math GCSEs (or 0.05σ on English).

Several prior studies also report the correlation between teacher FFT scores and student test scores. Our estimates from English secondary schools are in line with those other existing estimates. Studying teachers and younger students in the United States, but using similar data and regressions, prior papers report coefficients on FFT score of 0.08–0.09σ (Kane et al., 2011) and 0.05–0.11σ (Kane et al., 2013). The latter citation is from the large Methods of Effective Teaching (MET) Project, which included measuring teaching using other observation rubrics besides FFT, and generally the other rubrics also predicted test scores similarly (Kane & Staiger, 2012). A similar study of teachers and kindergartners in Ecuador found coefficients of 0.05–0.07σ for the CLASS rubric (Araujo et al., 2016). By contrast, (relatively) subjective ratings of teachers by school leaders are less consistently predictive student scores (Jacob & Lefgren, 2008; Rockoff & Speroni, 2010; Rockoff, Staiger, Kane & Taylor, 2012).

Our estimates are distinctive in two ways, even if they are similar in magnitude to prior estimates. First, the peer observers had relatively little training compared to prior studies. In prior studies, teachers were observed and rated by researchers or school administrators who receive

[16] Rothstein (2010) argues convincingly against the use of student fixed effects to study elementary school teachers, the most common setting in the literature. The requirement that schools use the same information is easily violated when the student fixed effects strategy uses observations over multiple years for a given student.

substantial training and are often tested for reliability before conducing evaluations.[17] Second, rubric ratings may depend on the instructional activities used during the observer's visit. For example, a rating of a teacher's "questioning and discussion techniques" may be more accurate or precise if the class spends more time in group discussion. We can control for instructional activities. Compare Table 7 columns 5 and 3. The coefficient on effectiveness ratings are quite similar whether or not we control for the mix of activities during the observed class.

Section 2.2 describes potential sources of measurement error in the FFT scores. If the sources of measurement error—sampling of visits, skew and ceiling effects, limited observer training—are uncorrelated with teachers' value-added to test scores, then our estimates—0.07σ for math, 0.04σ for English—will be biased too small. Indeed, the reliability of FFT scores is much less than one, so we should expect some classical attenuation bias. Alternatively, the measurement error could be correlated with value-added in ways that the estimates are biased too large. For example, peer observers might give higher FFT ratings to teachers they know make larger value-added contributions to student achievement, even if what the observers see during their visit does not warrant the higher ratings. While we cannot rule out these sources of bias, we note that our estimates are quite similar to prior estimates suggesting no substantial new bias in our setting.

3.5. Heterogeneity

Does the relationship between class activities and test scores change for different types of students? We find no evidence of heterogeneity. In Appendix Table A5 we re-estimate the specification in Table 7 column 1, but interact the instructional activity measures with the student's prior test score. None of the interactions are statistically significant, though the main effects of activities remain significant as they are in Table 7. For example, in English classes, student-peer interaction is effective but not more or less effective for students with lower prior achievement. Additionally, we extend the test by adding teacher fixed effects and again find no evidence of heterogeneity.

By contrast, the degree to which FFT instructional effectiveness ratings predict test scores does change with the student's prior achievement. As shown in Appendix Table A5, the correlation—between student GCSE scores and teacher effectiveness ratings—shrinks as the student's prior test scores rise.

4. Conclusion

This paper describes several results which contribute to answering the ongoing questions: What does effective teaching actually involve? What teaching practices matter for student achievement? We study teaching practices and student achievement in public (state) secondary schools in England.

Our primary focus is teachers' choices about how to allocate class time across different instructional activities. Classroom observers recorded how much class time was spent on different instructional activities—for example, "open discussion among children and teacher" and "use of white board by teacher." Observers simply recorded what activities were happening without judging the appropriateness or quality of the activity.

We find, in short, that teachers' choices of instructional activities predict their students' subsequent achievement scores. In math classes, for example, students score higher with teachers who give more time for individual practice. In the typical (average) math class, the teacher allocates "some of the time" to assessment and practice. In a class where the teacher allocates "most of the time" to practice and assessment, GCSE scores are 0.08σ higher than the typical class (σ = student test score standard deviations). For English exams, by contrast, more time working with classmates predicts higher scores.

Educators and researchers might well be skeptical that simple time use would predict student scores since teachers likely vary in how effectively they carry out different activities. Our data—with both time use and effectiveness measures—provides a rare opportunity to test that skeptic's hypothesis. When we control for instructional effectiveness ratings, class time use still predicts student achievement.

The practical implication of this paper is that students would likely gain (or lose) from changes in instructional activities even if teacher skills did not change. However, we caution against simply turning this one paper's specific activity groups into practice recommendations for teachers. As more evidence on this topic accumulates, practical steps will become clearer. In our data, 15–30 percent of the variation in time use is explained by the school, suggesting some potential for school-level interventions, but leaving most attention to teacher-level interventions.

We also caution against causal conclusions based solely on this paper. The apparent relationship between teachers' use of class time and student achievement may be caused by some unobserved teacher characteristic or behavior. In other words, there may yet be some further omitted variable bias in our estimates. However, our setting warrants stronger causal inferences than would be prudent in papers that only measure time use and lack any measure of teacher skill. Moreover, even if our estimates overstate the magnitude of the relationships, the direction and pattern of relationships should at least motivate further empirical analysis of class time use. We hope this paper will motivate a future field experiment to strengthen causal conclusions, and to test a practical intervention.

We also find that rubric-based ratings of instructional effectiveness also predict student achievement. A student assigned to a top-quartile teacher, as measured by effectiveness ratings, will score about 0.08σ higher than a similar student assigned to a bottom-quartile teacher. Classroom observations and rubrics are not new to schools or education researchers. Still, our data are novel in a few ways. Most notably, our observation data were collected by peer teachers—observer and observee were co-workers in the same school—and observers received little training—much less training than is often described as necessary for "valid" or "reliable" observations. In the end, we find peer ratings of instructional effectiveness predict at least as well as has been documented in other studies. Peer observation can be a feasible and effective approach to learning about differences in teaching, even with little additional training for observers.

One way to think about the magnitude our estimates is to ask what a 0.08σ improvement in GCSE scores would mean for a student's future. Indeed, GCSE scores are perhaps more relevant for students' futures, compared to tests at younger ages, because GCSEs come at the end of compulsory schooling and also inform college admissions. In a new analysis, Hodge, Little, & Weldon, 2021 estimate that a one standard deviation, 1σ, increase in average GCSE scores predicts about a 20 percent increase in lifetime earnings (NPV at age 16). Thus from 0.08σ we would predict a 1.6 percent increase in lifetime earnings, or about £7500 in present value at age 16.

Acknowledgements

We first thank the Nuffield Foundation for generous financial support of this analysis, and the Education Endowment Foundation for support of the original experiment which collected the data. Thanks for comments and advice to Anna Vignoles, Ellen Greaves, Hans Sievertsen, Julie Cohen, and Matthew Burgess, and to the Department for Education for access to the National Pupil Database.

Supplementary materials

Supplementary material associated with this article can be found, in

[17] Sometimes the raters are known as "peer evaluators" but "peer" refers to the fact that the rater had (recently) been a classroom teacher. The evaluator role is a distinct specialized job with substantial training.

the online version, at doi:10.1016/j.econedurev.2023.102405.

References

Angrist, J., Hull, P., Pathak, P.A. & Walters, C. (in-press). “Credible school value-added with undersubscribed school lotteries.” Review of Economics and Statistics.

Araujo, M. Caridad, Carneiro, Pedro, Cruz-Aguayo, Yyannú, & Schady, Norbert (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics, 131*(3), 1415–1453.

Aslam, Monazza, & Kingdon, Geeta (2011). What can teachers do to raise pupil achievement? *Economics of Education Review, 30*(3), 559–574.

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

Aucejo, Esteban, Coate, Patrick, Fruehwirth, Jane Cooley, Kelly, Sean, & Mozenter, Zachary (2020). *Match Effects in the Teacher Labor Market: Teacher Effectiveness and Classroom Composition*. Working paper.

Ballatore, RM, & Sestito, Paolo (2016). Dealing with student heterogeneity: Curriculum implementation strategies and student achievement. *Bank of Italy Temi di Discussione (Working Paper) No, 1081*.

Bloom, Nicholas, Eifert, Benn, Mahajan, Aprajit, McKenzie, David, & Roberts, John (2013). Does management matter? Evidence from India. *Quarterly Journal of Economics, 128*(1), 1–51.

Bloom, Nicholas, Lemos, Renata, Sadun, Raffaella, & Van Reenen, John (2015). Does management matter in schools? *Economic Journal, 125*(584), 647–674.

Bloom, Nicholas, & Van Reenen, John (2007). Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics, 122*(4), 1351–1408.

Burgess, Simon, Rawal, Shenila, & Taylor, Eric S. (2021). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. *Journal of Labor Economics, 39*(4), 1155–1186.

Campbell, Shanyce L., & Ronfeldt, Matthew (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267.

Chetty, Raj, Friedman, John N., & Rockoff, Jonah E. (2014). Measuring the impacts of teachers I: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2593–2632.

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

Duflo, Esther, Dupas, Pascaline, & Kremer, Michael (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739–1774.

Dwyer, Carol A., & Villegas, Ana M. (1993). *Guiding conceptions and assessment principles for the Praxis series: Professional assessments for beginning teachers*. Educational Testing Service Research Report RR-93-17.

Fitzpatrick, Maria D., Grissmer, David, & Hastedt, Sarah (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*, 269–279.

Gitomer, Drew, Bell, Courtney, Qi, Yi, McCaffrey, Daniel, Hamre, Bridget K., & Pianta, Robert C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–20.

Graham, Bryan S., Ridder, Geert, Thiemann, Petra, & Zamarro, Gema (2021). Teacher-to-classroom assignment and student achievement. *Working paper*.

Hanushek, Eric A., & Rivkin, Steven G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271.

Hayward, Hugh, Hunt, Emily, & Lord, Anthony (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 dfe-rr398a*. London: Department for Education.

Ho, Andrew D., & Kane, Thomas J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation.

Hodge, Louis, Little, Allan, Weldon, Matthew (2021). GCSE attainment and lifetime earnings. Department for Education Research Report DFE-RR1132.

ICPSR. (n.d.). “Measures of effective teaching: 3c - Base Data: Item-Level Observational Scores, 2009-2011 Variable Description and Frequencies.” ICPSR 34346.

Jackson, C. Kirabo, Rockoff, Jonah E., & Staiger, Douglas O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6*(1), 801–825.

Jacob, Brian A., & Lefgren, Lars (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.

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

Kane, Thomas J., & Staiger, Douglas O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER Working Paper no. 14607*.

Kane, Thomas J., & Staiger, Douglas O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, Thomas J., Taylor, Eric S., Tyler, John H., & Wooten, Amy L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587–613.

Kingdon, Geeta, Banerji, Rukmini, & Chaudhary, P. K. (2008). *SchoolTELLS survey of rural primary schools in bihar and uttar pradesh, 2007–08*. London: Institute of Education, University of London.

Kraft, Matthew A., & Gilmour, Allison F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249.

McIntosh, Steven. (2006). Further analysis of the returns to academic and vocational qualifications. *Oxford Bulletin of Economics and Statistics, 68*(2), 225–251.

Myford, Carol, Villegas, Ana Maria, Reynolds, Anne, Camp, Roberta, Danielson, Charlotte, Jones, Jacqueline, et al. (1994). Formative studies of Praxis III: Classroom performance assessments an overview. *Educational Testing Service Research Report RR-94-20*.

Prendergast, Canice. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7–63.

Rockoff, Jonah E., Jacob, Brian A., Kane, Thomas J., & Staiger, Douglas O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43–74.

Rockoff, Jonah E., & Speroni, Cecilia (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2), 261–266.

Rockoff, Jonah E., Staiger, Douglas O., Kane, Thomas J., & Taylor, Eric S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review, 102*(7), 3184–3213.

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

Sims, David P. (2008). Strategic responses to school accountability measures: It’s all in the timing. *Economics of Education Review, 27*, 58–68.

Slater, Helen, Davies, Neil M., & Burgess, Simon (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics, 74*(5), 629–645.

Syverson, Chad. (2011). What determines productivity? *Journal of Economic Literature, 49*(2), 326–365.

Taylor, Eric S. (2018). Skills, job tasks, and productivity in teaching: Evidence from a randomized trial of instruction practices. *Journal of Labor Economics, 36*(3), 711–742.

Taylor, Eric S., & Tyler, John H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628–3651.

Weisberg, Daniel, Sexton, Susan, Mulhern, Jennifer, Keeling, David, Schunck, Joan, Palcisco, Ann, et al. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York City: The New Teacher Project.