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, Simon.Burgess@bristol.ac.uk

[b]Oxford Partnership for Education Research and Analysis, 7 Queens Square, Lyndhurst Road, Ascot, Berkshire SL5 9FE, England, shenilarawal@aol.com

[c](corresponding author) Harvard University and NBER, Gutman Library 469, 6 Appian Way, Cambridge, MA 02138, eric_taylor@gse.harvard.edu

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.

JEL No. I21, J24

Keywords: Education production, Teacher performance

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 into 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,

[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.

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 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 et al. 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 related 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 et al. 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., Kane et al. 2011, Taylor 2018, Aucejo et al. 2020).

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 and Burgess 2011). Improvements in GCSE scores also predict future earnings and college going (Mcintosh 2006, Hayward, Hunt, and Lord 2014, Hodge, Little, and Weldon 2021).

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, Rockoff, and Staiger 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 et al. 2011, Kane and Staiger 2012, Kane et al. 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 and co-authors (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 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 (Mcintosh 2006, Hayward, Hunt, and Lord 2014, Hodge, Little, and Weldon 2021). 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 7,000 students who were taught by those teachers and for whom we have GCSE test scores. For math we have 5,211 students and 136 teachers, and for English 4,301

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) 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 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 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

[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.

[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.

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 2,700 classroom visits, where one observer scored one of her peer teachers. Visits typically lasted 15-20 minutes. 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 et al. 2011, Kane and Staiger 2012).

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, and Chaudhary 2008).

Teachers make quite different choices about how to spend class time. Figure 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 Figure A1 shows Figure 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 Figure 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 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 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 full rubric is provided in 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 1-12 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 behaviour" (mean 9.4) and lowest for "use of assessment" (mean 8.5). In general, teachers were rated more effective in classroom environment tasks than instruction tasks. "Use of assessment" also showed the largest differences in effectiveness

[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.

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 Figure 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

[6] In three prior studies in U.S. schools, ratings for the ten FFT items are correlated 0.72-0.88 (Kane and Staiger 2012, Ho and 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.

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 minutes. 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 Figure 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 (Weisberg et al. 2009, Kraft and Gilmour 2017), 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 Figure 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 Figure 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 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 and 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 and 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

[7] The variation is likely due in part to using a 12 point scale, instead of the conventional 4 point scale. Appendix Figure A3 shows a histogram of the same data as Figure 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.

treatment improves value-added (Taylor and Tyler 2012, Burgess, Rawal, and Taylor 2022).[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, and Kremer 2011, Ballatore and Sestito 2017, Aucejo et al. 2020, Graham et al. 2021). Alternatively, the judgements of classroom observers may be influenced by the students in the class during the visit (Campbell and Rondfeldt 2018).

However, in this paper's setting, we find little evidence of a relationship between observable student 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.

[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 and Villegas 1993, Myford et al. 1994, Danielson 2007).

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:

$$A_{ijs} = T_j\delta + X_{ijs}\beta + \lambda_s + \varepsilon_{ijs}, \quad (1)$$

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.

[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.

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

$$A_{ijks} = T_{jk}\delta + X_{ijs}\beta + \lambda_s + \theta_k + \epsilon_{ijks}, \quad (2)$$

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.

[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.

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] 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, Rockoff, and Staiger 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 (Sims 2008, Fitzpatrick, Grissmer, and Hastedt 2011, Aucejo and Romano 2016).

[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.

[15] Slater, Davies, and Burgess (2011) 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 (2011) 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 estimates 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.

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., Kane et al. 2011, Taylor 2018, Aucejo et al. 2020).

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 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 test taken five years prior to the GCSE tests, not the immediate prior school year. Well known evidence—from Kane and Staiger (2008), Chetty, Friedman, and Rockoff (2014), 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.

[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.

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 Figure 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 et al. (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 Figure A1 shows that, on average, teachers do not differentiate activities by subject when being watched by their peers.

Finally, as detailed in Appendix C, our results are robust 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 and 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 and Lefgren 2008, Rockoff and Speroni 2010, Rockoff et al. 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 substantial training and are often tested for reliability before conducing evaluations.[17] Second, rubric ratings may

[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.

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 question: What does effective teaching actually involve? Or 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, and 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 £7,500 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.

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*.

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

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

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

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

Ballatore, Rosario Maria, and Paolo Sestito. (2016). "Dealing with student heterogeneity: curriculum implementation strategies and student achievement." Bank of Italy Temi di Discussione (Working Paper) No, 1081.

Bloom, Nicholas, and John Van Reenen. (2007). "Measuring and Explaining Management Practices Across Firms and Countries." *Quarterly Journal of Economics, 122* (4), 1351–1408.

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

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

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. (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., and Matthew Ronfeldt. (2018). "Observational Evaluation of Teachers: Measuring More Than We Bargained For?" *American Educational Research Journal, 55* (6), 1233–1267.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. (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*. Association for Supervision and Curriculum Development, Alexandria, VA.

Dwyer, Carol A., and Ana M. Villegas. (1993). *Guiding Conceptions and Assessment Principles for the Praxis Series: Professional Assessments for Beginning Teachers.* Educational Testing Service Research Report RR-93-17.

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

Fitzpatrick, Maria D., David Grissmer, and Sarah Hastedt. (2011). "What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment." *Economics of Education Review 30*, 269- 279.

Garet, Michael S., Andrew J. Wayne, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske. (2017). *The Impact of Providing Performance Feedback to Teachers and Principals. NCEE 2018-4001.* US Department of Education.

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

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

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

Hayward, Hugh, Emily Hunt, and Anthony Lord. (2014). *The economic value of key intermediate qualifications: estimating the returns and lifetime productivity gains to GCSEs, A levels and apprenticeships*. Department for Education Report DFE-RR398A. London: Department for Education.

Ho, Andrew D., and Thomas J. Kane. (2013). *The Reliability of Classroom Observations by School Personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

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

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

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

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

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

Kane, Thomas J., and Douglas O. Staiger. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle, WA: Bill & Melinda Gates Foundation.

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

Kingdon, Geeta, Rukmini Banerji, and P. K. Chaudhary. (2008). *SchoolTELLS Survey of rural primary schools in Bihar and Uttar Pradesh, 2007–08.* London: Institute of Education, University of London.

Kraft, Matthew A. and Allison F. Gilmour. (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, Ana Maria Villegas, Anne Reynolds, Roberta Camp, Charlotte Danielson, Jacqueline Jones, Joan Knapp, Penny Lehman, Ellen Mandinach, Lori Morris, Alice Sims-Gunzenhauser, and Barbara Sjostrom. (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., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. (2011). "Can you recognize an effective teacher when you recruit one?" *Education Finance and Policy, 6* (1), 43-74.

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

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. (2012). "Information and employee evaluation: Evidence from a randomized intervention in public schools." *American Economic Review, 102* (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, Neil M. Davies, and Simon Burgess. (2012). "Do teachers matter? Measuring the variation in teacher effectiveness in England." *Oxford Bulletin of Economics and Statistics, 74* (5), 629-645.

Syverson, Chad. (2011). "What determines productivity?" *Journal of Economic Literature, 49* (2), 326-265.

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., and John H. Tyler. (2012). "The effect of evaluation on teacher performance." *American Economic Review, 102* (7), 3628-3651.

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

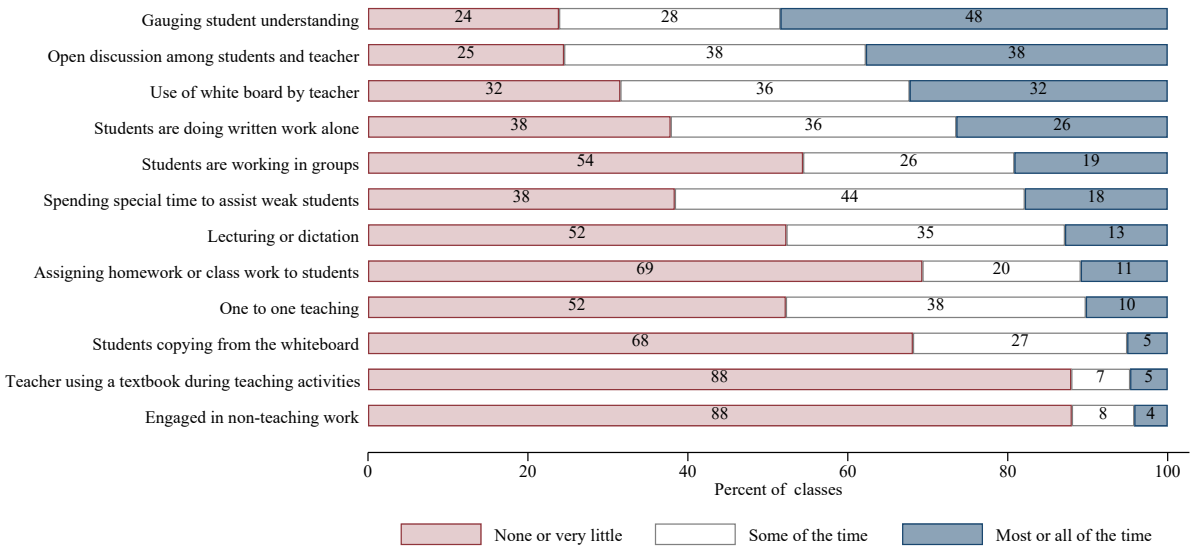
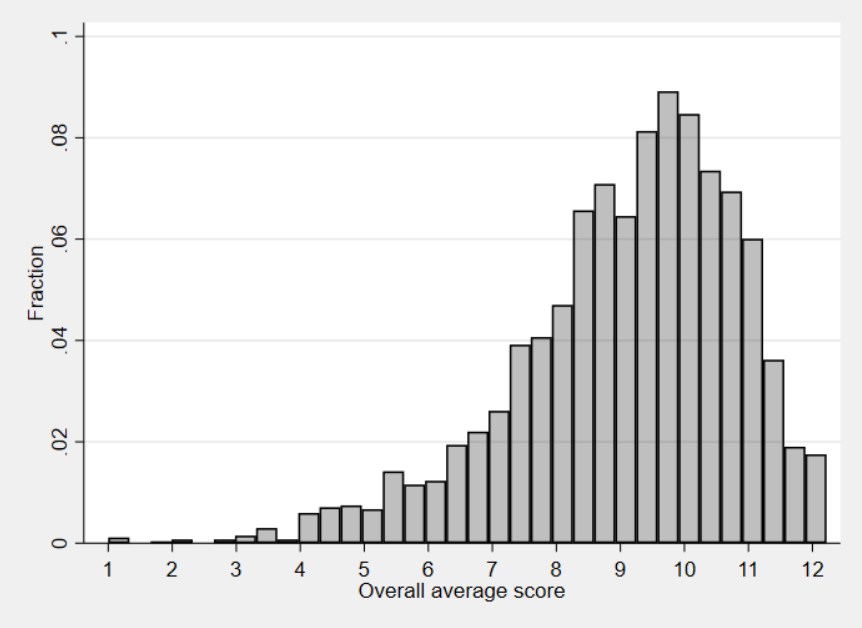


Figure 1—Frequency of instructional activities

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 2,687 observations, each the visit of a peer observer k to the class of teacher j.

(A) Average of item scores



(B) Item-level scores

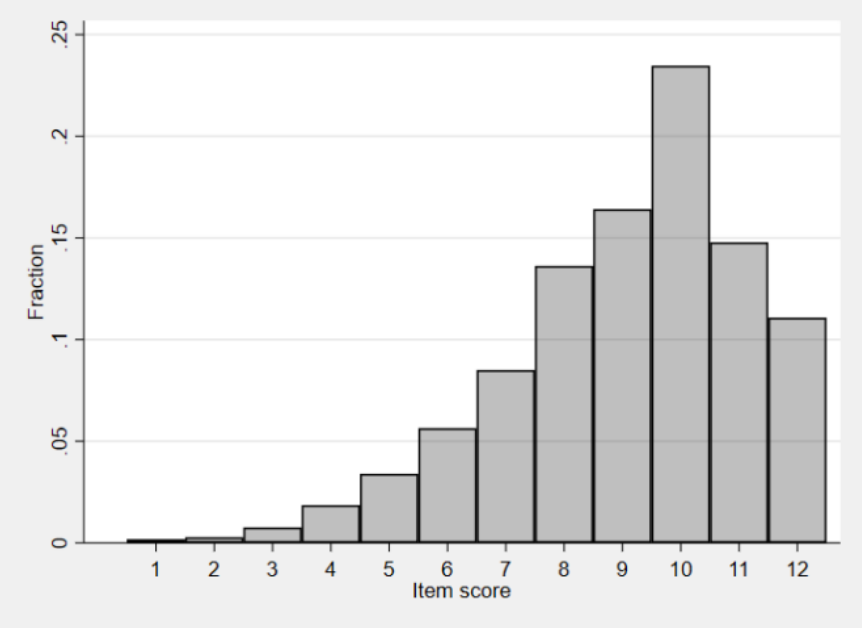


Figure 2—Distribution of instructional effectiveness ratings

Note: Panel A shows a histogram of 2,687 average scores. Each of the 2,687 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 23,047 item-level scores recorded across the rubric's ten items.

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, Banerji, and Chaudhary 2008). The grouping of activities in the right hand column is described in the text.

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 behaviour against those standards. The teacher response to student misbehaviour is consistent, proportionate, appropriate and respects the students' dignity.
1.e Organising 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.

Table 2 (cont.)—Rubric standards and associated description of "Effective"

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 rigour. 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 *Framework for Teaching* (Danielson 2007).

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.

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 2,687 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.

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	1.36	1.06
					(0.71)	(0.72)	(0.66)
Student peer interaction	0.17	1			1.68	1.71	1.66
					(0.93)	(0.91)	(0.94)
Personalized instruction	0.21	0.31	1		1.44	1.52	1.34
					(0.92)	(0.91)	(0.92)
Practice and assessment	0.37	0.28	0.35	1	1.58	1.73	1.38
					(0.91)	(0.86)	(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 2,687 observations. Each of the 2,687 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	Math	English
	(1)	(2)	(3)
Overall average	9.09	9.15	9.00
	(1.75)	(1.80)	(1.69)
Classroom environment average	9.27	9.35	9.17
	(1.84)	(1.88)	(1.78)
1a. Creating an environment of respect and rapport	9.32	9.35	9.28
	(2.04)	(2.09)	(1.97)
1b. Establishing a culture for learning	9.20	9.25	9.13
	(2.01)	(2.04)	(1.96)
1c. Managing classroom procedures	9.24	9.31	9.14
	(2.04)	(2.06)	(2.01)
1d. Managing student behaviour	9.41	9.42	9.41
	(2.05)	(2.12)	(1.96)
1e. Organising physical space	9.13	9.29	8.87
	(2.18)	(2.14)	(2.23)
Instruction average	8.90	8.94	8.86
	(1.83)	(1.87)	(1.77)
2a. Communicating with students	9.29	9.31	9.25
	(1.91)	(1.95)	(1.85)
2b. Using questioning and discussion techniques	8.77	8.80	8.72
	(2.17)	(2.16)	(2.18)
2c. Engaging students in learning	8.99	9.03	8.93
	(2.00)	(2.09)	(1.86)
2d. Use of assessment	8.50	8.53	8.46
	(2.21)	(2.19)	(2.23)
2e. Demonstrating flexibility and responsiveness	8.83	8.78	8.90
	(2.05)	(2.08)	(2.01)

Note: Means and standard deviations (in parentheses), using a sample of 2,687 observations in column 1. Each of the 2,687 observations is the visit of a peer observer k to the class of teacher j. The samples for columns 2 and 3 are 1,510 and 1,177 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.

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 9,512 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

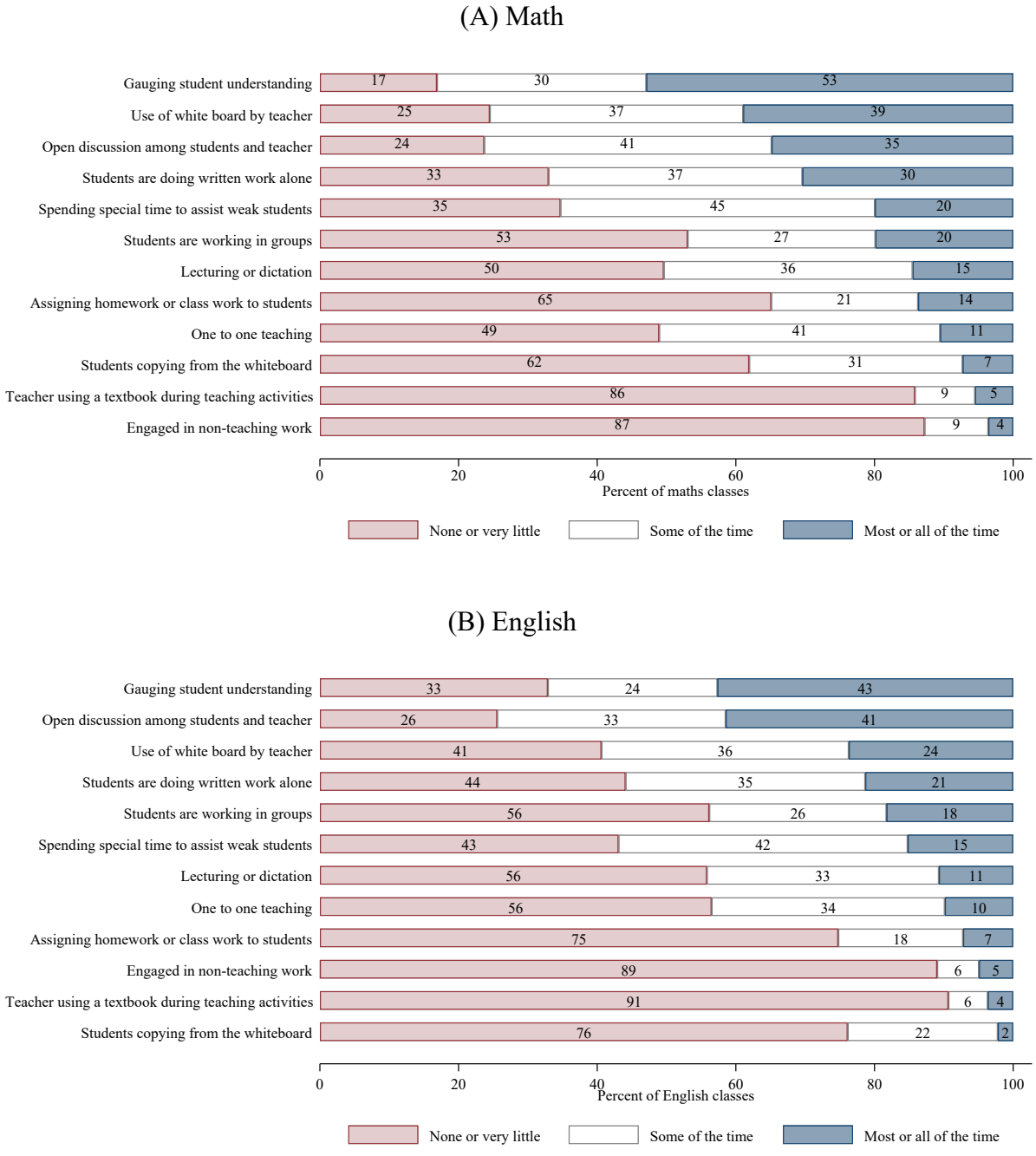
Teachers' use of class time and student achievement

Online appendix

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

March 2023

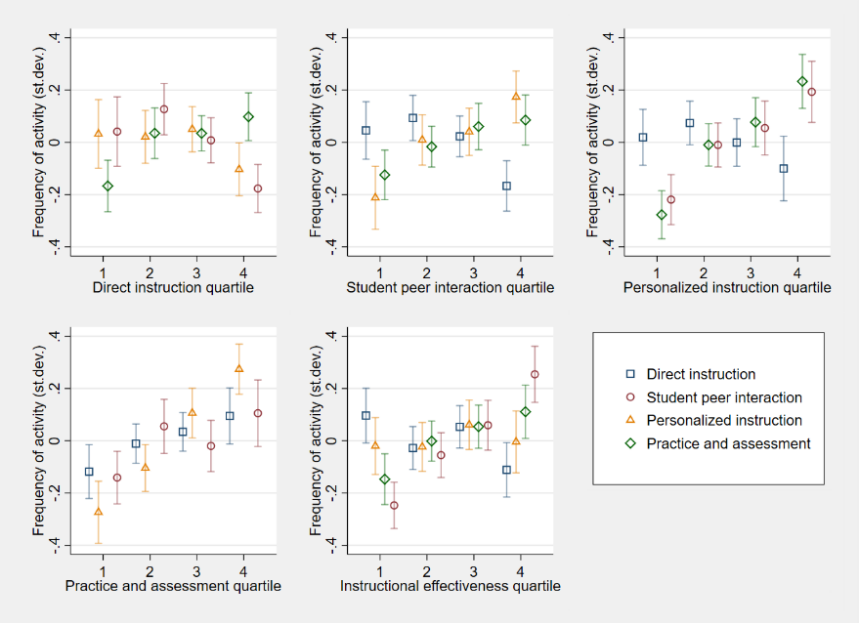
Appendix A: Additional figures and tables



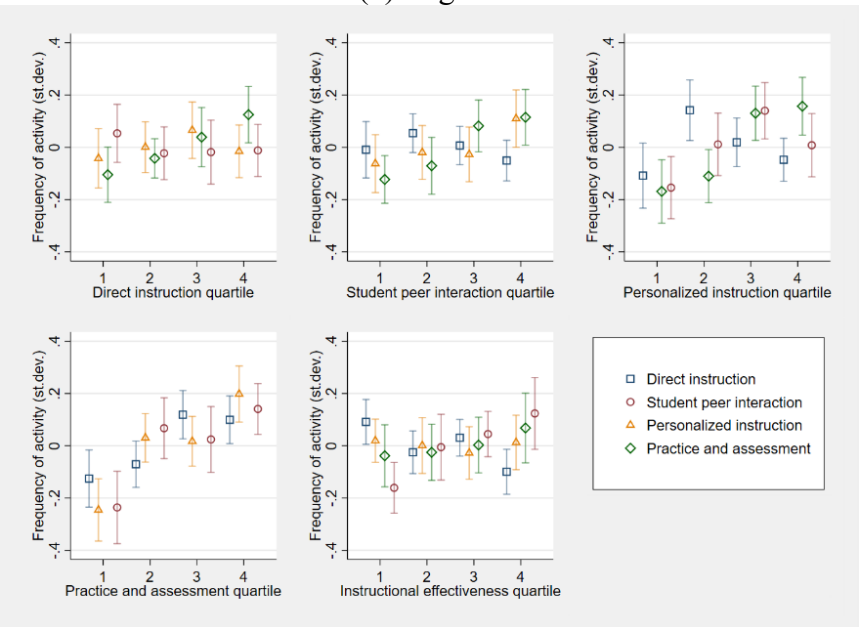
Appendix Figure A1—Frequency of instructional activities, by subject

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 grey (middle) bar is the "some of the time." Panel A is for math classes, and the proportions are of 1,510 observations, each the visit of a peer observer k to the class of a math teacher j. Panel B is for English classes and based on 1,177 observations.

(a) Math

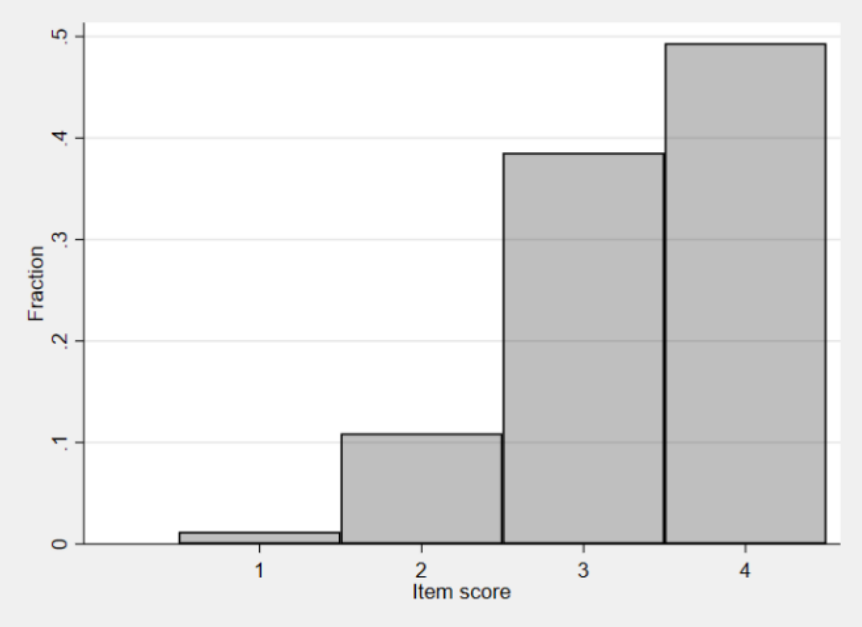


(b) English



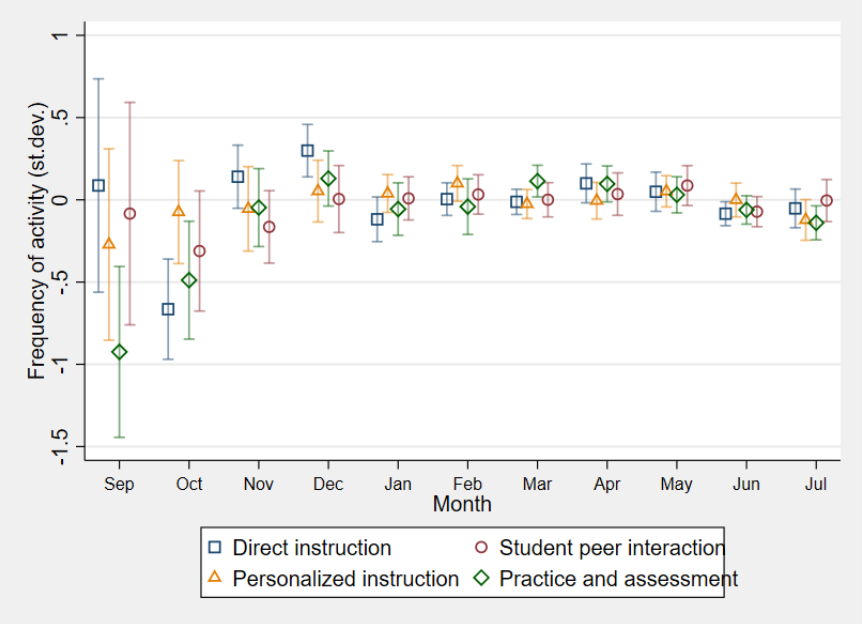
Appendix Figure A2—Co-occurrence of instructional activities, by subject

Note: In the top-left plot, for a given subject, the x-axis divides observations into quartiles of "direct instruction," where 4 is teachers who allocate the most class time to direct instruction. The y-axis measures the frequency (in standard deviation units) of the other three activities: student peer interaction, personalized instruction, and practice and assessment. The markers are the mean of the activity measure, net of observer fixed effects. The vertical bars show the 95 percent confidence using cluster (teacher *j*) corrected standard errors. All the other plots are constructed the same way. In the bottom-middle plot the x-axis quartiles are based on average FFT rating. Panel A for math has 1,510 observations, and Panel B for English has 1,177 observations.



Appendix Figure A3—Distribution of instructional effectiveness ratings, item-level scores, 4-point scale

Note: Histogram of the 23,047 item-level scores recorded across the rubric's ten items. This figure shows the same item-level data as Figure 2 panel B, except that the 12-point scale for scores has been collapsed to a 4-point scale: scores 1-3 in panel A become a scores of 1 in panel B, 4-6 become 2, 7-9 become 3, and 10-12 become 4.



Appendix Figure A4—Frequency of instructional activities by month

Note: This figure reports the frequency of instructional activities by month of observation. Each of the 2,687 observations is the visit of a peer observer k to the class of teacher j. The x-axis divides those 2,687 observations into the month during which the visit occurred. The y-axis measures the frequency of each activity (in standard deviation units). The markers are the mean of the activity measure, net of observer fixed effects. The vertical bars show the 95 percent confidence using cluster (teacher j) corrected standard errors.

Appendix Table A1—Correlations among instructional effectiveness ratings

		(1a)	(1b)	(1c)	(1d)	(1e)	(2a)	(2b)	(2c)	(2d)	(2e)
1a.	Creating an environment of respect and rapport	1									
1b.	Establishing a culture for learning	0.86	1								
1c.	Managing classroom procedures	0.79	0.81	1							
1d.	Managing student behaviour	0.79	0.80	0.82	1						
1e.	Organising physical space	0.68	0.68	0.69	0.67	1					
2a.	Communicating with students	0.74	0.76	0.72	0.71	0.65	1				
2b.	Using questioning and discussion techniques	0.65	0.67	0.63	0.61	0.55	0.75	1			
2c.	Engaging students in learning	0.73	0.77	0.71	0.72	0.61	0.79	0.76	1		
2d.	Use of assessment	0.62	0.65	0.63	0.59	0.58	0.66	0.68	0.72	1	
2e.	Demonstrating flexibility and responsiveness	0.70	0.71	0.66	0.66	0.62	0.75	0.74	0.77	0.73	1

Note: Correlations of rubric-based instructional effectiveness ratings among ten tasks, using a sample of 2,687observations. Each of the 2,687observations is the visit of a peer observer *k* to the class of teacher *j*. Observers rated effectiveness on a 1-12 scale: 1-3 ineffective, 4-6 basic, 7-9 effective, and 10-12 highly effective.

Appendix Table A2—Principal components of instructional effectiveness ratings

	Original units		Net of observer fixed effects	
	Component		Component	
	1	2	1	2
	(1)	(2)	(3)	(4)
Weight in component				
1a. Creating an environment of respect and rapport	0.33	-0.27	0.33	-0.28
1b. Establishing a culture for learning	0.33	-0.23	0.34	-0.25
1c. Managing classroom procedures	0.32	-0.34	0.32	-0.37
1d. Managing student behaviour	0.32	-0.37	0.32	-0.38
1e. Organising physical space	0.29	-0.30	0.27	-0.20
2a. Communicating with students	0.33	0.16	0.33	0.18
2b. Using questioning and discussion techniques	0.30	0.44	0.30	0.44
2c. Engaging students in learning	0.33	0.22	0.33	0.21
2d. Use of assessment	0.30	0.39	0.29	0.41
2e. Demonstrating flexibility and responsiveness	0.31	0.33	0.31	0.32
Eigenvalue	7.43	0.69	6.60	0.78
Proportion of variation explained	0.74	0.07	0.66	0.08

Note: Principal component analysis of rubric-based instructional effectiveness ratings among ten tasks, using a sample of 2,687observations. Each of the 2,687observations is the visit of a peer observer k to the class of teacher j. Observers rated effectiveness on a 1-12 scale: 1-3 ineffective, 4-6 basic, 7-9 effective, and 10-12 highly effective. The main body of the table reports the component loadings, where loadings are the weights given to each item (rows) in calculating the score for a given component (columns). Columns 1-2 report components 1-2 using unadjusted effectiveness ratings, as recorded by observer k. Columns 3-4 report components 1-2 using effectiveness ratings net of observer fixed effects. For columns 3-4, before the principal component analysis, we first calculate observer k's mean for each item and subtract that mean from all scores k assigned for that item.

Appendix Table A3—Student characteristics, activities, and ratings

	Prior test score	Class st.dev. prior test score	Female	Month of birth	Ever free school meals	IDACI score
	(1)	(2)	(3)	(4)	(5)	(6)
	(A) Instructional activities, math					
Direct instruction	0.002	0.006	-0.004	0.000	0.002	-0.006**
	(0.039)	(0.018)	(0.009)	(0.042)	(0.008)	(0.002)
Student peer interaction	-0.033	0.017	-0.002	0.010	-0.012	0.002
	(0.045)	(0.020)	(0.009)	(0.044)	(0.011)	(0.002)
Personalized instruction	-0.026	-0.014	0.014+	-0.051	0.003	0.001
	(0.029)	(0.019)	(0.007)	(0.039)	(0.007)	(0.002)
Practice and assessment	-0.029	0.011	0.007	0.110**	0.008	-0.001
	(0.031)	(0.019)	(0.007)	(0.037)	(0.009)	(0.003)
	(B) Instructional effectiveness, math					
Instructional effectiveness	0.057+	0.001	0.005	0.016	-0.008	-0.001
	(0.034)	(0.016)	(0.011)	(0.036)	(0.009)	(0.002)
	(C) Instructional activities, English					
Direct instruction	0.022	-0.008	-0.012	-0.060	-0.010	-0.006+
	(0.043)	(0.019)	(0.010)	(0.050)	(0.011)	(0.003)
Student peer interaction	0.018	0.026	-0.013	-0.036	-0.011	-0.005
	(0.035)	(0.017)	(0.008)	(0.044)	(0.012)	(0.003)
Personalized instruction	-0.024	-0.014	0.005	0.034	0.009	0.003
	(0.031)	(0.015)	(0.009)	(0.035)	(0.009)	(0.002)
Practice and assessment	-0.069+	0.025	-0.026**	-0.043	0.007	0.002
	(0.040)	(0.019)	(0.009)	(0.051)	(0.008)	(0.003)
	(D) Instructional effectiveness, English					
Instructional effectiveness	0.073*	-0.006	0.012	0.033	-0.006	-0.005+
	(0.036)	(0.016)	(0.009)	(0.047)	(0.010)	(0.003)

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 9,512 student-by-subject observations. Each column within each panel reports results from a separate regression. All estimation details are the same as for Table 7 with these exceptions: The dependent variable—described in each column header—is a baseline characteristic of student i or student i's classmates for subject s. The only controls are observer fixed effects, and time on "non-teaching work" for panels A and C.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A4—Instructional activities and student achievement scores, accounting for the timing of observations

	(1)	(2)	(3)	(4)	(5)	(6)
			(A) Math			
Instructional activities						
Direct instruction	0.012	0.008	0.023	0.020		
	(0.016)	(0.015)	(0.015)	(0.015)		
Student peer interaction	0.020	0.024*	0.002	0.005		
	(0.013)	(0.012)	(0.014)	(0.013)		
Personalized instruction	0.004	0.000	0.003	-0.000		
	(0.019)	(0.019)	(0.019)	(0.019)		
Practice and assessment	0.068**	0.063**	0.047*	0.041*		
	(0.019)	(0.019)	(0.019)	(0.019)		
Instructional effectiveness			0.070**	0.070**	0.077**	0.078**
			(0.018)	(0.018)	(0.017)	(0.017)
			(B) English			
Instructional activities						
Direct instruction	-0.018	-0.016	-0.009	-0.008		
	(0.020)	(0.020)	(0.019)	(0.020)		
Student peer interaction	0.053**	0.052**	0.043**	0.043**		
	(0.016)	(0.016)	(0.016)	(0.016)		
Personalized instruction	-0.021	-0.019	-0.026+	-0.023+		
	(0.013)	(0.014)	(0.014)	(0.014)		
Practice and assessment	-0.024	-0.026+	-0.030+	-0.033*		
	(0.016)	(0.014)	(0.017)	(0.015)		
Instructional effectiveness			0.039+	0.039*	0.040*	0.038*
			(0.021)	(0.019)	(0.019)	(0.018)
Student covariates	√	√	√	√	√	√
Month of observation 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 9,512 student-by-subject observations. Columns 1, 3, and 5 in this table simply reproduce columns 1, 3, and 5 from Table 7. For columns 2, 4, and 6, the details of estimation are identical to 1, 3, and 5, respectively, with one exception. In the even numbered columns we add additional controls in the form of indicators for month of observation. If an observer, k, visited an observee, j, in more than one month, then more than one indicator will be equal to one.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix Table A5—Differences by students' prior test scores

	Pooled		Math		English	
	(1)	(2)	(3)	(4)	(5)	(6)
(A) Instructional effectiveness rating						
Instructional effectiveness	0.060**		0.074**		0.039*	
	(0.014)		(0.017)		(0.019)	
Instructional effectiveness	-0.027*	-0.023*	-0.030*	-0.028+	-0.016	-0.008
* prior test score	(0.011)	(0.012)	(0.015)	(0.016)	(0.015)	(0.014)
(B) Instructional activities						
Direct instruction	-0.003		0.013		-0.017	
	(0.012)		(0.015)		(0.020)	
Direct instruction	-0.008	-0.002	-0.025	-0.012	0.001	0.002
* prior test score	(0.011)	(0.011)	(0.016)	(0.016)	(0.015)	(0.012)
Student peer interaction	0.034**		0.019		0.053**	
	(0.010)		(0.014)		(0.016)	
Student peer interaction	-0.001	-0.006	0.002	0.001	-0.001	-0.006
* prior test score	(0.013)	(0.014)	(0.020)	(0.022)	(0.013)	(0.012)
Personalized instruction	-0.004		0.006		-0.022	
	(0.012)		(0.019)		(0.014)	
Personalized instruction	-0.005	-0.010	-0.008	-0.009	-0.001	-0.006
* prior test score	(0.013)	(0.014)	(0.021)	(0.023)	(0.016)	(0.015)
Practice and assessment	0.020		0.070**		-0.024	
	(0.014)		(0.019)		(0.016)	
Practice and assessment	-0.009	-0.007	-0.025	-0.026	0.005	0.004
* prior test score	(0.014)	(0.015)	(0.023)	(0.025)	(0.014)	(0.014)
Teacher 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 9,512 student-by-subject observations. Panel A: All estimation details are the same as for Table 7 column 5 with these exceptions: We interact teacher j's instructional effectiveness rating with student i's prior test score in subject s, recall $j = j(is)$. In even numbered columns, we also include teacher j fixed effects. Panel B: All estimation details are the same as for Table 7 column 1 except that we add interactions similar to panel A.
+ indicates p<0.10, * 0.05, and ** 0.01

Appendix B: Instructional effectiveness rubric

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.

Note: Adapted from *Framework for Teaching* (Danielson 2007).

Appendix C: Instructional activities, principal components

Our main analysis divides the twelve instructional activities into four mutually exclusive and exhaustive categories (as in Tables 1, 5, and 7). An alternative characterization is the principal components of the twelve activity items. Appendix Table C1 shows the principal components details, and Appendix Table C2 uses the component scores to predict student GCSE scores similar to Table 7.

There are tradeoffs between the two approaches. The complex weights in the principal components approach capture (potentially) more-nuanced latent dimensions of teaching practice. The disadvantage is that the principal components are more difficult to describe in words. Thus, we caution that substantive conclusions should not depend on the "correctness" of short-hand labels we attach to principal components. By contrast, the four simple groups are straightforward to understand, but use less of the variation in how activities are correlated.

Each principal component score is a weighted average of the twelve activities. The weights are shown in Appendix Table C1. The first five principal components together explain just over half of the variation in the instructional activities data; each of the five individually explain 9-13 percent. By construction, the principal component scores are uncorrelated with each other.[1]

We use the following labels for the principal components scores, though others may choose alternative labels. (1) "Student-teacher interaction." This component score is increasing in the amount of class time where teacher and students are interacting.[2] (2) "Smaller groups vs. whole class." This score is increasing in individual and small group activities, and decreasing in whole class activities. (3) "Practice vs. instruction." This score is increasing in student

[1] Before estimating the principal components we first rescale the activities item data. Observers record the frequency of an activity on a 0-4 scale with 0 "none" of the time to 4 "full time" during the observation. To rescale we divide each of the twelve items by the sum of the items. Thus, the rescaled items measure the proportion of all activity recorded by the observer.
[2] For expositional purposes we reverse the sign of components (1), (2), and (3).

assessment and practice, and decreasing in instruction, especially individualized instruction. (4) "Group vs. individual work." This score is increasing in time where students are interacting with classmates, and decreasing in time where students are working alone or one-on-one with teacher. (5) "Teacher guided learning." This score is increasing in gauging understanding and assisting weak students, and use of the white-board, and decreasing in open discussion, children working alone and one-way lecturing.

The principal components of instructional activities do predict student test scores, as shown in Appendix Table C2. For English GCSEs, the fourth principal component stands out from the others. The coefficient for component four is 0.05σ, while all other estimates are less than 0.01σ and far from statistically significant. Our short-hand label for component four is "group vs. individual work"; it 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.

Both the principal components approach, in Appendix Table C2, and the simpler grouping of activities, in Table 7, end in a similar substantive conclusion for predicting English test scores. Both emphasize activities where students interact with their classmates. But these are not two independent tests, and the general similarity should not be surprising. Both approaches combine activities based on the same correlation matrix in Table 4. Still, the principal component weights, shown in Appendix Table C1, are quite different from the approach in Table 7 which weights items equally but in mutually exclusive and exhaustive groups.

Math GCSE scores are predicted, first, by the third principal component. Our short-hand label for this component is "practice vs. instruction." This pattern is consistent with the simple groups approach in Table 7. In fact, (i) the third component "practice vs. instruction" is correlated 0.81 with (ii) the difference "practice and assessment" minus "personalized instruction" in the simple groups.

The fifth principal component also predicts student math scores. The estimated coefficients for the third and fifth components are similar in magnitude and precision. The third explains 11 percent of the variation in the activity item data, but the fifth explains nearly as much at 8 percent. The fifth component suggests some potential additional insight is lurking in the activities data. However, the fifth principal component is difficult to describe in words. Our best attempt at a parsimonious description is “teacher guided learning.”

To summarize, the instructional activities teachers choose to use in their classes partly explain student achievement growth. In math, students score higher when more time is devoted to student practice and assessment. In English, students score higher when they spend more time working and talking with their classmates. These patterns are robust to how we go about combining activities into groups or components. Nor do our conclusions rely on the short-hand descriptions we use for groups or components.

Appendix Table C1—Principal components of instructional activities

	Component				
	1	2	3	4	5
	(1)	(2)	(3)	(4)	(4)
Weight in component					
1. Open discussion among children and teacher	-0.43	0.12	0.05	0.27	-0.45
2. Children are working in groups	-0.11	-0.26	-0.02	0.61	0.14
3. One to one teaching	0.04	-0.29	0.55	-0.15	-0.22
4. Spending special time to assist weak students	-0.10	-0.20	0.51	-0.22	0.42
5. Children are doing written work alone	0.20	-0.26	-0.26	-0.54	-0.35
6. Gauging student understanding	-0.32	-0.19	-0.54	-0.09	0.19
7. Assigning homework or class work to children	0.38	-0.08	-0.22	0.12	0.22
8. Lecturing or dictation	0.14	0.51	0.06	0.04	-0.34
9. Children copying from the whiteboard	0.29	0.48	0.09	0.03	0.14
10. Use of white board by teacher	-0.24	0.42	-0.04	-0.31	0.43
11. Using a textbook during teaching activities	0.39	-0.06	0.07	0.27	0.15
12. Engaged in non-teaching work	0.44	-0.12	-0.14	0.01	-0.07
Eigenvalue	1.55	1.48	1.34	1.27	1.06
Proportion of variation explained	0.13	0.12	0.11	0.11	0.09

Note: Principal component analysis of class time use among twelve instructional activities, using a sample of 2,687 observations. Each of the 2,687 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 the principal component analysis, we first rescaled the data, dividing each of the twelve 0-4 item scores by the sum of the item scores for the observation. The main body of the table reports the component loadings, where loadings are the weights given to each item (rows) in calculating the score for a given component (columns).

Appendix Table C2—Instructional activities and student achievement scores, principal components approach

	(1)	(2)	(3)	(4)
(A) Math				
Instructional activities				
1. "Student-teacher interaction"	0.022	0.013	-0.002	0.006
More time where teacher and students are interacting	(0.020)	(0.009)	(0.020)	(0.008)
2. "Smaller groups vs. whole class"	0.019	0.007	0.008	0.006
More time in individual and small group activities, less time in whole class activities	(0.015)	(0.006)	(0.015)	(0.007)
3. "Practice vs. instruction"	0.039*	0.008	0.028+	0.005
More time on student assessment and practice, less time on instruction, especially individualized instruction	(0.016)	(0.006)	(0.016)	(0.006)
4. "Group vs. individual work"	-0.014	0.001	-0.020	-0.001
More time where students are interacting with classmates, less time working alone or one-on-one with teacher	(0.014)	(0.008)	(0.013)	(0.008)
5. "Teacher guided learning"	0.048**	0.019**	0.041**	0.017*
More time using the whiteboard and assisting students, less Time solo working and one-way lecturing.	(0.014)	(0.007)	(0.013)	(0.007)
Instructional effectiveness			0.071**	0.022*
			(0.018)	(0.009)
(B) English				
Instructional activities				
1. "Student-teacher interaction"	-0.009	0.012	-0.018	0.008
	(0.013)	(0.008)	(0.012)	(0.007)
2. "Smaller groups vs. whole class"	0.007	0.006	0.001	0.003
	(0.012)	(0.007)	(0.012)	(0.007)
3. "Practice vs. instruction"	0.006	0.007	0.002	0.005
	(0.011)	(0.006)	(0.011)	(0.006)
4. "Group vs. individual work"	0.047**	0.022**	0.045**	0.023**
	(0.011)	(0.006)	(0.012)	(0.006)
5. "Teacher guided learning"	0.003	0.002	0.003	-0.000
	(0.014)	(0.009)	(0.014)	(0.009)
Instructional effectiveness			0.044*	0.025*
			(0.021)	(0.011)
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 9,512 student-by-subject observations. All estimation details are the same as for Table 7 with one exception: We replace the four instructional activities groups with five principal component scores.
+ indicates p<0.10, * 0.05, and ** 0.01