Economics of Education Review 82 (2021) 102091

Contents lists available at ScienceDirect

Economics of Education Review

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

Who teaches the teachers? A RCT of peer-to-peer observation and feedback in 181 schools☆

Richard Murphy [*,a], Felix Weinhardt [b], Gill Wyness [c]

[a] University of Texas at Austin, NBER, IZA, CESifo and Centre for Economic Performance
[b] European University Viadrina, DIW Berlin, IZA, CESifo and Centre for Economic Performance
[c] UCL Institute of Education and Centre for Economic Performance

ARTICLE INFO

JEL classification:
I21
I28
M53

Keywords:
Education
Teachers
RCT
Peer mentoring

ABSTRACT

This paper evaluates a widely used, low stakes, teacher peer-to-peer observation and feedback program under Randomized Control Trial (RCT) conditions. Half of 181 volunteer primary schools in England were randomly selected to participate in a two-year program in which three fourth and fifth grade teachers observed each other. We find that two cohorts of students taught by treated teachers perform no better on externally graded national tests compared to business as usual. However this masks large heterogeneity; in small schools, where there is only one class per grade, we find negative impacts of the training (0.1-0.18SD), whereas we find positive impacts in larger schools (0.06-0.17SD). We outline and explore potential mechanisms for this and conclude that centralised one-size-fits-all teacher training interventions may be harmful.

1. Introduction

It is well established that teachers are the most important in-school factor in determining student outcomes (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Thus, the fact that there is huge variation in teacher quality (Hanushek & Rivkin, 2010) is a perennial problem for education policy-makers concerned with student test scores, and their consequences for earnings and welfare (Barro, 2001; Chetty, Friedman, & Rockoff, 2014; Hanushek & Woessmann, 2015). One obvious course of action to improve student outcomes would be to hire better teachers; however, many studies have concluded that teacher effectiveness is very difficult to predict from teacher characteristics (Aaronson, Barrow, & Sander, 2007; Kane, Rockoff, & Staiger, 2008), reducing the viability of this solution. An alternative would be to simply dismiss poorly performing teachers (Chetty et al., 2014; Hanushek & Rivkin, 2010), but this too is a challenge given the administrative burden required, difficulties with finding replacements and lack of good measures of teacher effectiveness available to school principals (Jacob, Rockoff, Taylor, Lindy, & Rosen, 2016; Rothstein, 2015).

Consequently, a potentially powerful strategy for policy-makers concerned with improving educational outcomes would be to improve the quality of the stock of existing teachers either through incentives or teacher training programs. Research in this area has tended to focus on the former, with a number of studies evaluating the use of performance related pay as a means to improve teacher productivity (Goodman & Turner, 2010; Lavy, 2009; Muralidharan & Sundararaman, 2011; Neal, 2011; Springer et al., 2011). However, these studies have had mixed results, calling into question the effectiveness of performance related pay as a 'magic bullet' to improve educational outcomes in developed countries. An alternative means of improving teacher performance on-the-job, which has received much attention in recent years, is through observation based teacher training programs.[1] Taylor and Tyler (2012) and Burgess, Rawal, and Taylor (2021) both find positive evidence on the effectiveness of one particular type of teacher development - teacher feedback. Recent work has also found evidence of teacher co-worker spillovers from job transitions (Jackson & Bruegmann, 2009)

☆ We thank Stephen Machin, Chris Karbownik, Anna Raute, Stephen Rivkin and Eric Taylor for valuable feedback, as well as participants of the Bonn/BRIC Economics of Education Conference, the Mannheim labour seminar and IWAEE. We thank the UK Department for Education for access to the English student census data under DR160317.03 and the Education Endowment Foundation for funding. Weinhardt gratefully acknowledges financial support by the German Research Foundation through CRC TRR 190 (project number 280092119). All errors are our own.

* Corresponding author.

[1] Non-observation based methods of teacher training have failed improve teacher effectiveness in experimental (Garet et al., 2011; 2010) or quasi-experimental settings (Harris & Sass, 2011; Jacob & Lefgren, 2008). The exception is (Angrist & Lavy, 2001) which does find a positive effect.

https://doi.org/10.1016/j.econedurev.2021.102091
Received 10 April 2020; Received in revised form 29 January 2021; Accepted 2 February 2021
Available online 17 March 2021

as well as from targeted teacher training interventions. (Papay, Taylor, Tyler, & Laski, 2016)[2]

In this paper, we estimate the causal effect of teacher observation and feedback on student outcomes by implementing a Randomized Control Trial (RCT) in 181 primary schools in England with above average levels of poverty.[3] The treatment allocated to 89 of these schools over a two year period is one of the most popular teacher observation programs in the world, Lesson Study. Lesson Study is a teacher peer-to-peer learning approach found in more than 50 countries and increasingly practiced in the U.S. (Akiba & Wilkinson, 2016; Lewis, Perry, Hurd, & O'Connell, 2006; Perry & Lewis, 2009; Robinson, 2015).[4] It consists of a group of teachers planning and observing each others lessons, and providing feedback as a means to constructively improve their teaching. In our setting, fourth and fifth grade teachers work in groups of three, with the first teacher being observed three times by her two peers over the course of a month. This process is then repeated for the remaining two teachers over the course of the academic year. As the program was implemented for two academic years this resulted in a total of eighteen lesson observations.[5] To ensure structured feedback and implementation, all participating teachers received five training days held by educational experts on teaching mentoring. Our outcomes of interest come from national, compulsory, high stakes, externally marked academic tests conducted at the end of primary school, one year after the intervention ends. We access these test scores, and other pupil characteristics, from detailed administrative data linked to our program.

Overall, we find no evidence that teacher peer observation and feedback increases pupil performance compared to "business as usual" in the classroom. The relatively high power of the trial means that for the first cohort we can reject effect sizes larger than 0.07σ on student performance across all subjects, and effect sizes larger than 0.05σ in reading and writing. Therefore, at most, the program is half as effective as reducing elementary school classes by seven pupils (Krueger & Whitmore, 2001). However, the overall null finding masks consistently large heterogeneity related to school size. As will be explained in more detail in Section 3, the original research design was to sample small schools (with a single class per grade) to avoid teacher or student selection, but this sample was expanded for recruitment purposes to include large schools (with more than one class per grade). Due to the size of the trial, we have sufficient power to examine the importance of this by splitting our sample into small (intended) schools, and large (unintended) schools. While only the intended sample was pre-registered, by definition this deviation to our intended sample was effectively registered (and defined) as the unintended sample.

The program has negative effects on student performance in small schools and positive effects in larger schools. These impacts are larger in magnitude for the second cohort of students that had twice as much exposure to the program. The negative impact on small schools increases from -0.10σ to -0.19σ, in contrast the impact in larger schools increased from improving student outcomes by 0.07σ to 0.17σ. This is consistent with the program being worse than the status quo in smaller schools but better in larger schools.

We explore the mechanisms behind the clear differences in the impact of the program between small and larger schools. These can be broadly categorized into 'compositional' 'selection', 'matching' and 'disruption' effects. First, it may be the case that larger schools are systematically different in a number of other dimensions which makes the program more effective. While school size is correlated with other factors, such as student composition, achievement, or quality of school leadership, we find no evidence that these factors relate to the gains from the program. A second possible explanation is that in schools with multiple classes per grade, the principal could select the teachers/students that would gain the most from the training. We find no evidence of selection of students within schools that had the choice.

Third, the gains in larger schools could also be driven by matching effects. One argument for why peer-to-peer learning is effective is that it provides an opportunity for less effective teachers to learn from others. An assumption of this mechanism is that there is enough heterogeneity of teacher quality among the participants to guarantee meaningful information flows, a situation which is more likely in larger schools than smaller ones. This hypothesis is difficult to test, since we do not have data on teachers themselves. However, we cannot rule it out, and it is backed up by the literature, e.g. Papay, Tyler, and Taylor (2018) find positive co-worker effects among pairs of teachers who were purposefully paired up based on previous effectiveness measures.

Finally, our results are also consistent with the intervention being more disruptive in smaller schools. An accompanying process evaluation of the program from during the trial found evidence indicating that small schools faced more organizational challenges from the introduction and perpetual implementation of the program.[6] Moreover, in smaller schools, teachers do not have the option to opt in and so may be less committed to the program. Our result that the negative impact of the program increases over time, implies that a one-time disruption effect is unlikely to be driving the results, and that small schools face more disruption in both years of the program.

We conclude that the large differences in the effectiveness are not primarily driven by other school characteristics, student selection or one-time disruption effects (although we cannot definitively rule these mechanisms out), while matching effects - more plausible in larger schools, and disruption effects - more likely in smaller schools, are consistent with our findings.

Our conclusion, and a key contribution of this paper, is that a "one size fits all" approach to teacher peer-to-peer learning may not achieve the desired results. Though our results provide suggestive evidence that matching of teachers is important to the success of peer-to-peer learning, ultimately we cannot rule out other mechanisms for the different effects by school size. Nevertheless, our results provide more encouragement for this type of teacher feedback programme in larger schools.

There are a number of reasons why we might expect peer-to-peer mentoring to be an effective form of teacher training. Unlike many other professions, teachers do not interact with their peers in the classroom. Thus, classroom observations offer an opportunity for teachers to see, and be seen in action. Feedback on their observed performance could thus provide teachers with new detailed information on their performance in the classroom. Given that teachers have been shown to be "motivated agents" (Dixit, 2002), this could result in improved planning and preparation and subsequently better performance (Steinberg & Sartain, 2015). Perhaps unsurprisingly then, many schools already carry out some form of peer observation, albeit with little instruction or consistency (Weisberg et al., 2009), making them difficult to evaluate empirically. Indeed some teachers in our study reported having had experience of using classroom observation in the past,

[2] Full literature review at the end of this section.

[3] Defined by share of students eligible for Free School Meals being greater than then national average of 18 percent

[4] The majority of districts in Florida have mandated the use of Lesson Study (Akiba, Murata, Howard, & Wilkinson, 2019).

[5] Three teachers, each being observed three times per year, over two years.

[6] A full process evaluation took place alongside this quantitative study, including observation of the teacher training, interviews with staff involved in the treatment, and analysis of data on control schools' use of peer observation approaches. This qualitative evaluation was based on visits to 10 schools in two of the three implementation regions, and interviews with 19 staff and senior managers. Follow up interviews were conducted by telephone and email with five expert teachers in five schools, and information on progress provided by four other schools. The process evaluation was led by an independent team from the National Institute of Economic and Social Research, who provided a report in May 2015. Full details can be found at Murphy, Weinhardt, and Wyness (2017).

for appraisal or development purposes. These tended to be more informal, short observations (e.g. for 10 minutes) often without accompanying feedback. For example, in describing a previous experience, one school pointed out that "the process as a whole was not sufficiently structured to identify areas of improvement with sufficient accuracy and detail. By contrast, the program we study is well-established and structured, involving an intensive training program.

Moreover, testing the impact of teacher observation, and teacher training in general, on pupil outcomes is an empirical challenge due to non-random selection of teachers (and students) into training. Our trial is large-scale, with 543 teachers teaching 13,000 students, over two cohorts in all subjects, across 181 primary schools in England. Despite having strict experimental conditions, our experiment is conducted within schools, in a manner which could easily be replicated or taken to scale. Thus, we capture the impact of teacher observation and feedback in a "real-world" setting.

This study is directly related to the small but growing literature on observation based teacher training and student outcomes. In contrast to our findings, these papers typically find positive significant effects, but the interventions differ in potentially important ways which help to explain the differences and which contribute to our understanding of the mechanisms for success.

Jackson and Bruegmann (2009) show evidence of teachers learning from co-workers by studying job transitions of teachers. They find that newly arriving effective teachers, measured by their students' value added, improve the effectiveness of their co-workers. Moreover, these effects persist even when teachers move on to teach in different schools. In contrast to this informal learning setting, we study a widely used structured training program where teachers are compelled to observe and provide feedback to their existing peers.

Taylor and Tyler (2012) use the as-good-as random roll-out of a teacher observation program across middle schools in Cincinnati[7] finding positive effects of teacher peer observation. The program involved each teacher having three unannounced observations by external experts, and one by the school principal. After each observation teachers were provided with formal written feedback and grades, which had consequences, including impact on promotions, tenure, and potential non-renewal of the teacher's contract. The study finds that the students of teachers who have been evaluated improve their maths scores by 0.11σ in the year after the evaluation, and about 0.16σ two years later. These effects are in line with our estimates for large primary schools, which are similar in size to American middle schools. In addition to the difference in school size, the nature of the Lesson Study program differs from the Cincinnati intervention in two key ways. First, it does not involve teacher incentives; as it is intended to facilitate free and open discussion between the teachers no formal scoring or consequences are associated with the observations. Second, observations are conducted by the teachers peers rather than external experts or principals. Both of these factors would make the program cheaper and easier to expand to scale, but, as our results imply, the Lesson Study program could not guarantee the presence of a high quality observer (heterogeneity in teacher effectiveness), whereas this is more plausible in the Cincinnati setting, where observers were external. This is an inherent limitation of peer-to-peer training: the lack of external experts limits the variation in knowlege within a training group to what already exists within a school.

Also relevant is evidence from Papay et al. (2016) who study teacher peer-to-peer training in an intervention that paired high- and low-performing teachers together, with the goal to improve the low-performing teachers skills through learning from a higher performer. They find that students in classrooms of low-performing treated teachers score 0.12σ higher. Again, this lends credence to our hypothesis that the ability to identify high quality teachers is important for the effectiveness of these types of programs.

Most recently, Burgess et al. (2021) conduct a low stakes peer observation experiment in 82 high schools in England. Teachers in treatment schools were randomly selected to be either observers, observees or both. They show positive effects of the treatment on student achievement, for both pupils of observer and observee teachers. While this study is similar to ours in the sense of teachers being observed on multiple occasions, with low stakes, there are some key differences. First, our program is based on the well-established 'Lesson Study' program which is a peer to peer learning program, whereas theirs is explicitly an evaluation program (albeit a low-stakes one). Second, the Burgess et al study takes place in high schools; these are around 4 times larger, with 5 times as many teachers than the primary schools of our setting[8] therefore adding weight to our finding of positive impacts may be more likely in larger schools.

As well as contributing to this small and growing literature on teacher peer to peer learning, our paper provides the first experimental evaluation of a teacher observation program that is in use throughout the world. Through obtaining an informative null result for all schools in our sample, and being able to reject effect sizes larger than 0.09σ, our results show that a blanket approach to teacher observation and feedback cannot solve the policy maker's problem of poorly performing teachers, and caution against centralized and prescriptive policies for teacher training. At the same time, we document positive effects in large schools in which there is greater likelihood of there being a high quality teacher present for others to learn from. Exploring teacher-level heterogeneity in program-effects, and understanding whether such programmes can be effective in smaller schools, are important routes for future work on this subject.

The remainder of the paper proceeds as follows: Section 2 provides further details about the intervention. In Section 3 we describe the data used in the analysis, with the RCT design described in Section 4. Results are presented in Section 5, and potential mechanisms are explored in Section 6. In the final section we discuss the effect sizes in comparison to other estimates from the education literature, and broader conclusions.

2. Institutional context and data

2.1. Institutional context

In England, pupils attend primary school from age 4/5 to 10/11, taking them from Reception through to Year 6, where grades are called Years or Year Groups. For the purpose of clarity we will refer to them as grades. The educational curriculum is organised around Key Stages, where Key Stage 2 incorporates grades 3, 4, 5, and 6. This trial was conducted in the last three years of Key Stage 2, which are the last years of primary education, and at the end of which pupils are evaluated.

2.2. Student census data

We use administrative data that are available for all students in state-education in England from the National Pupil Database (NPD). Pupils take national Key Stage 1 tests in grade 2 at age 6/7 and the Key Stage 2-test at the end of primary school at age 10/11. From now on we refer to these tests as age-7 and age-11 tests.

The administrative age-7 tests serve as the baseline measure of student achievement. Students are assessed in math and reading and are

[7] Papay et al. (2018) currently have an ongoing teacher observation RCT in the field with an end date of 2020. A pilot study by Steinberg and Sartain (2015) evaluates the Chicago Excellence in Teaching Project (EIP) in which teachers are observed by their principal during a lesson, followed by a feedback session as well as more formal ratings, finds no significant effect.

[8] In 2014 there were 13.3 FTE teachers per primary school, and 64.1 FTE teachers per secondary school (for Education, 2014; UK, 2013)

Calendar Year / School Year	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015	2015/2016
Year 2 (Age-7 tests - Controls)	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4	Cohort 5	Cohort 6
Year 3	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4	Cohort 5
Year 4	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3	Cohort 4
Year 5	Cohort -2	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**	Cohort 3
Year 6 (Age-11 tests - Outcomes)	Cohort -3	Cohort -2	Cohort -1	Cohort 0	**Cohort 1**	**Cohort 2**

Fig. 1. Timeline of intervention Notes: Red square shows treatment period and cohorts.

graded by their teacher, with test scores taking values between three and 27. Since these national tests are available for all students, we use the mean of reading and maths achievement level as our measure of initial student ability.

The age-11 tests examine the students ability in four different areas: maths, reading, Spelling Punctuation and Grammar (SPAG) and science. We use the first three of these, which are externally marked on a 100 point scale, and used to assign students a national achievement level (between 2-5). We percentalised the raw score at the national subject-cohort level to ensure comparability across subjects and years.

The use of the administrative test score data has several key benefits. First, this data is available for all students and schools with no attrition from the data in the treatment or control groups. Second, we have a comparable measure of the students' achievement prior to the intervention. Third, after control schools were informed that they were not selected into treatment we did not need to contact them again, or conduct any testing in these schools. Fourth, the information is available for previous cohorts of students, allowing us to test for balance in outcomes for prior cohorts. Finally, no additional testing was required to assess the impact of this program, thus the tests are not tailored to the intervention. Indeed, it has been shown that performance in these national age-11 exams is a strong predictor of later outcomes, including wages (DfE, 2013). This means we can estimate effects of the program on an outcome measure which has known benefits.

3. Details of intervention

3.1. Lesson Study

Lesson Study is a peer-to-peer observation and feedback program with a long history of use in Japan, and now increasingly used in the US and worldwide.[9] The key element of Lesson Study is teacher observation and feedback: teachers work in small groups to plan lessons that address shared teaching and learning goals, observe each others lessons and give feedback. In our setting, teachers were advised that the focus of their observation and feedback would be student learning in literacy (including tackling underachievement in reading, guided reading, groupwork for literacy) and numeracy (including tackling underachievement in mathematics, effective feedback in mathematics, and errors and misconceptions in multiplication and division). The leadership from the treated schools attended a senior leader conference at the beginning of the programme. This provided them with more details for how the training would be rolled out in their schools.

In our setting, teachers within a school form a group of three (known as a "learning tripod"), with one of the three selected as the "expert teacher". The implementation of the program starts with an initial group meeting where the three teachers plan the order in which they are to be observed and which lessons will be observed. The first teacher then teaches her three "research lessons" observed by the other two teachers. During these classes, the observing teachers do not interact with students or the teacher but remain solely in their observing role. After each class the group meets to discuss the lesson and plan the next in terms of content, structure and delivery. Over the course of the academic year there are three cycles of the program with each teacher taking their turn to be observed. Of course, it is likely that pupils would be aware of the presence of the observing teacher and this could impact their behaviour and future outcomes in and of itself. The impact of this is ambiguous - it could be that students are distracted by the presence of the observer, negatively impacting their performance. On the other hand, their behaviour could improve as a result of the in-class student teacher ratio increasing. Either way, this should be considered part of the treatment.

The lack of formal scoring highlights that the program's intention is to provide a space for non-judgemental discussion in the school day, rather than a formal evaluation program incorporating consequences or incentives, such as that considered by Taylor and Tyler (2012) and, to an extent Burgess et al. (2021). The process evaluation confirmed this, emphasising that teachers welcomed the input from peer observation, particularly with its emphasis on support, rather than performance management. For example, one teacher commented that the approach made it possible to convey to an under-performing teacher what they need to do to improve in a more supportive way. Teachers reported positively on the experience of sharing practice with teacher colleagues, shared planning, and identifying complementary skills.

Nevertheless, Lesson Study is a structured program and so the teachers received five in-depth training days to prepare them for the program. This was conducted by experts in the program and included information on the ethos, protocols and practice of Lesson Study. Four of the five training days occurred during the first year. The fifth training day, at the beginning of the second year, was focused on optimising feedback and sustaining the program through its second year. Thus, while the program lasted for two years - and potentially changed teacher practice and student learning for much longer - the training for the intervention was heavily concentrated in the first year. Although short, the training may have also had an impact on the behaviour of the teachers. This could be positive - by providing additional motivation, or negative - by taking teachers out of the classroom for five days, or replacing their teacher training. Thus our parameter of interest will reflect the combination of teacher observation and the training required.

Since teacher improvement through observation could affect pupil performance in many areas, we estimate the impact of the program on all tested subjects at the end of primary school. These are maths, reading

[9] For example Lesson Study Alliance (http://www.lsalliance.org/) helps US teachers, mainly based in Chicago, use Lesson Study; Fernandez, Cannon, and Chokshi (2003) study a US Japan lesson study collaboration; Perry and Lewis (2009) describe the use of Lesson Study in a medium-sized California K-8 school district, and Akiba and Wilkinson (2016) in Florida.

and Spelling Punctuation and Grammar (SPAG). Our pre-specified main outcome of interest is the students mean performance in reading and maths.

The trial was pre-registered with the American Economic Association's registry for RCTs and a detailed statistical analysis plan was approved before we had access to the administrative student outcomes data.[10] The program was delivered independently of this impact evaluation by a team at Edge Hill University with support from external consultants.[11]

3.2. Timing of intervention

The trial of the program took place in state primary schools in England[12] during the 2013/14-2015/16 academic years. Figure 1 shows the affected cohorts in calendar years and the target in terms of academic years. In this paper, we analyse effects on age-11 outcomes for two cohorts, which were affected by one (cohort 1) or two years (cohort 2) of this intervention, both measured one year after the end of the intervention. We estimate the impact on each cohort seperately to not impose any functional form on how the second cohort (who are more exposed to the program and taught by teachers with more experience of the program) are impacted differently.

3.3. Recruitment and teacher selection

In order to maximize statistical power and achieve the smallest Minimum Detectable Effect Size (MDES), we wanted to recruit as many schools as funding would allow. Ultimately, we secured funding for 80 schools to be treated. Instead of using a roll-out RCT (which was originally proposed by the funder) where the control schools would eventually receive the treatment limiting us to 40 in the treatment group, and 40 in the control group, our design provided no training to the control schools at any point during or after the trial had ended. This doubled the size of trial allowing there to be 80 treatment schools and 80 control schools, which provides a MDES of 0.1σ (see Appendix Figure A.1).[13] This is in line with a recent review of education field experiments by Lortie-Forgues and Inglis (2019) who find the average effect size to be 0.06σ.

The target population for this study are state primary schools in England with above average Free School Meal eligibility (FSM), which stood at 18 percent at the time of randomisation in 2013 (DfE, 2016), and one class per cohort. The former was a requirement of the Education Endowment Foundation, the funder of the trial, the latter was to keep selection of teachers and students into the program within the randomly selected schools to a minimum.

The project developers were asked to recruit such primary schools in three regions in England in which they had capacity to deliver the program.[14] The regions were the South West, East Midlands and North West. Each region contains a number of Local Authorities (LAs) that are responsible for the running of schools in that area.[15] In order to recruit schools the developers first had to obtain the approval of the relevant LAs. In the end, we recruited schools from 18 LAs (see Appendix 1 for the complete list). The aim of the recruitment was to eventually have 160 schools participate in the study. This total was determined by baseline power calculations (see Appendix Figure A.1) to capture effect sizes of 0.1σ

Ultimately, 182 schools agreed to participate in the trial by sending back signed expression of interests. One of these schools one was ineligible, as it was a new school and would not have a cohort of students taking the age-11 tests during the evaluation period and therefore was excluded. This left 181 schools to be randomized into treatment or control status as described in the randomization section below, and a MDES of 0.09σ.

All of these schools signed an agreement to grant us access to their NPD data prior to randomization. After randomization, the 89 schools selected for treatment additionally signed a Memorandum of Understanding which stated the responsibilities of the schools, practitioners, and the evaluation team.[16]

Overall, the recruitment phase led to 6,436 participating students in the first cohort and 6,298 in the second cohort, for which we have administrative age-11 outcomes available.[17]

The recruiting team had difficulty recruiting schools meeting the initial target population requirements of having one class per grade and above national average FSM. In the end, half (91) of the recruited schools had more than one class per grade and 79 had below 19 percent FSM student population.

The motivation for the focus on one-class-per grade schools was to limit teacher selection. This is because schools were very free to choose which teachers were to be involved in the intervention, and who would be the expert teacher (though all schools chose teachers with some subject expertise in English or maths as the expert). The only restriction we placed on schools was that two of the chosen teachers should be teaching grades 4 and 5. In schools with only one class per grade, this means that both of their teachers from years 4 and 5 had to be selected, with one additional teacher joining from another grade. In contrast, larger schools could meet the requirement of having two teachers from the fourth and fifth grades without choosing all of their teachers of these grades.[18]

While our data does not allow us to identify the teachers chosen for the intervention, there is a lower possibility of teacher or student selection in the smaller schools, where there is no practical choice over participating classes in the target grades. Indeed, the process evaluation pointed out the relation of teacher selection with school size: "In smaller schools with two, or even one teacher for each year group, there was little choice over the team composition and selection was not therefore strategic".[19]

To explore the issues relating to teacher selection we split the sample

[10] The AEA trial registration number is 1779, for details see: http://www.socialscienceregistry.org/trials/1779. The statistical pre-analysis plan can be accessed here: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_4-Lesson_study_SAP.pdf

[11] See https://everychildcounts.edgehill.ac.uk/special-projects/lesson-study/ for more details.

[12] 93 percent attend state primary schools in England (DfE, 2015)

[13] The implementor advised us to power our experiment to detect sizes of between 0.4-07σ.

[14] Hypothetically schools could have been recruited from all over the country, however we wanted to minimize travel costs for teachers who had to travel to the training days.

[15] These are considerably larger than school districts in America with 152 currently operating in England. Unlike American school districts they have no power to raise finances to pay for school facilities; funding for education is provided to LAs from the central government who then allocate it across schools.

[16] In order to motivate schools to participate in this teacher development program we had to ensure that they did not perceive this intervention as useful for teacher assessment. One implication of this is that we could not collect and merge-in teacher-level information.

[17] There are 362 students (5 percent) for which the full set of demographics and attainment data was not available. This was approximately evenly split between treatment (172) and control groups (190).

[18] Some of the smallest schools had mixed-age classes, and so one teacher may have taught both Year 4 and Year 5. These very small schools had no choice regarding the involvement of teachers for years 4 and 5.

[19] See Murphy et al. (2017, p. 36)

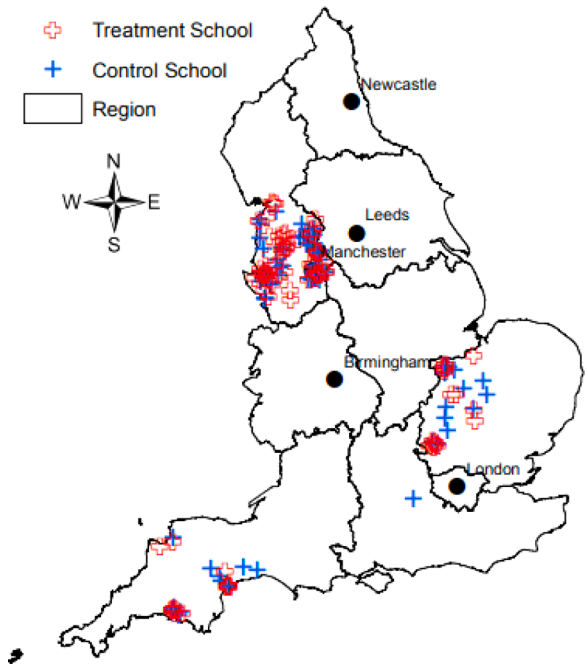

Fig. 2. Treatment and control schools

Table 1
National and local representativeness of sample

Variable	(1) National	(2) Local	(3) Sample (all)	(4) Intended	(5) Unintended
Age-7 Test	15.709	15.643	15.269	15.530	15.429
	[3.917]	[3.897]	[3.787]	[3.756]	[3.715]
Age-11 Maths Level	3.047	3.036	3.015	3.025	3.043
	[0.986]	[0.990]	[0.944]	[0.954]	[0.936]
Age-11 English Level	2.988	2.970	2.909	2.936	2.295
	[0.974]	[0.981]	[0.953]	[0.956]	[0.941]
Share Free School Meals	0.181	0.192	0.223	0.232	0.199
	[0.385]	[0.394]	[0.416]	[0.422]	[0.399]
Share Female	0.489	0.492	0.499	0.499	0.495
	[0.500]	[0.500]	[0.500]	[0.500]	[0.500]
Share Special Edu. Needs	0.137	0.142	0.160	0.168	0.159
	[0.344]	[0.349]	[0.367]	[0.374]	[0.366]
English Second Languge (ESL)	0.156	0.121	0.169	0.057	0.191
	[0.363]	[0.327]	[0.375]	[0.393]	[0.366]
Minority	0.218	0.166	0.208	0.081	0.242
	[0.413]	[0.372]	[0.406]	[0.272]	[0.428]
N	554,768	69,346	6,372	2,472	3,286

Notes: This table shows baseline characteristics for a pre-treatment cohort sitting the age-11 tests in maths and english in 2011. Note that for this cohort age-11 test scores were only avaialable to us in levels at the time of the randomsation, so this is what we report here. Column 1 includes all students of that cohort, column 2 only students in the same Local Authority and column 3 students of the schools that were part of the trial. Column 4 is for schools where both schools in the randomization pair have cohort sizes below 31 in our treatment years - the intended schools - and column 5 for pairs of schools which both have larger cohort sizes. Standard deviations of variables shown in square parenthesis.

of schools into small schools (our intended sample) and large schools (unintended sample). While this part of the analysis was not pre-registered, it is far from chance. Indeed, when we originally registered the trial, we specified and expected to sample only one class per grade schools precisely to avoid the issue of teacher selection. We define a school to be small if it has 30 of fewer students in each grade. The reason for this is that 30 is the maximum class size for these Year Groups. Schools are defined to be large if they have over 30 students in each grade we observe.

3.4. Representativeness

Figure 2 shows the geographical position of the schools in our sample, the red crosses denote schools of the treatment group and the blue crosses of the control group. We can see the schools come from three regions with the exception of one school in the south east of England. Table 1 shows how the schools within our sample compare with all schools nationwide and within the participating authorities, using information from students who completed their age-11 tests in 2011, three years prior to the intervention. In line with the recruitment strategy, pupils in our sample are more likely to have FSM (22 percent) than pupils nationally (18 percent) or within their LA (19 percent). Pupils are more likely to be minority than in within their LA (17 percent), though are representative of the national mix. They are marginally more likely to have English as a Second Language (ESL) (17 percent) than pupils nationally (16 percent) or within their LA (12 percent). The students are more likely to possess a statement of Special Educational Needs (16 percent) than pupils nationally (14 percent) or locally (14 percent). As may be expected given our research design of targeting schools with above average proportions of low income students, the average attainment at age-7 in these schools is 0.45 levels lower than schools nationally. For the outcomes, age-11 tests, the students perform worse in English by 0.08 levels (0.08σ), but achieve comparably to schools nationally or locally in maths (0.03σ). The proportion female and the cohort size are similar among our sample and schools locally and nationally. Taken as a whole, the schools in our sample contain slightly more disadvantaged students than an average school, and have a better value added in maths, but they are not distinctly different and therefore we have confidence in the external validity of the trial.

3.5. Randomization and compliance

Randomization: We performed a pairwise stratified randomization of schools by LA with the aim of balancing the randomization at LA level (i.e. the pairing of schools for randomization was conducted within each LA). This was to ensure there were equal numbers of treated and control schools within each local authority and that they would be balanced in terms of unobservable local characteristics.

In order to pair similar schools within LAs we computed an index score using principal component analysis based on school level characteristics. These characteristics were taken from before the intervention in 2011, and consisted of the school level average maths and reading levels of students in their age-11 tests and the share of students eligible for FSM.

As described above, 181 schools agreed to participate in the trial and were eligible for the program. Treatment status was initally only allocated to schools for which we could construct an index score (8 schools had no age-11 test scores in 2011) and schools that did not operate as part of pair-franchise (6 schools). This left 167 schools of which 83 schools were assigned to treatment and 84 were assigned to control.[20] When the 83 selected schools were informed that they would be treated,

[20] The randomization procedure is explained in more detail in Murphy et al. (2017). We only had funding for the treatment to be implemented in 80 schools. We assumed that not all the 83 assigned schools would go through with the experiment

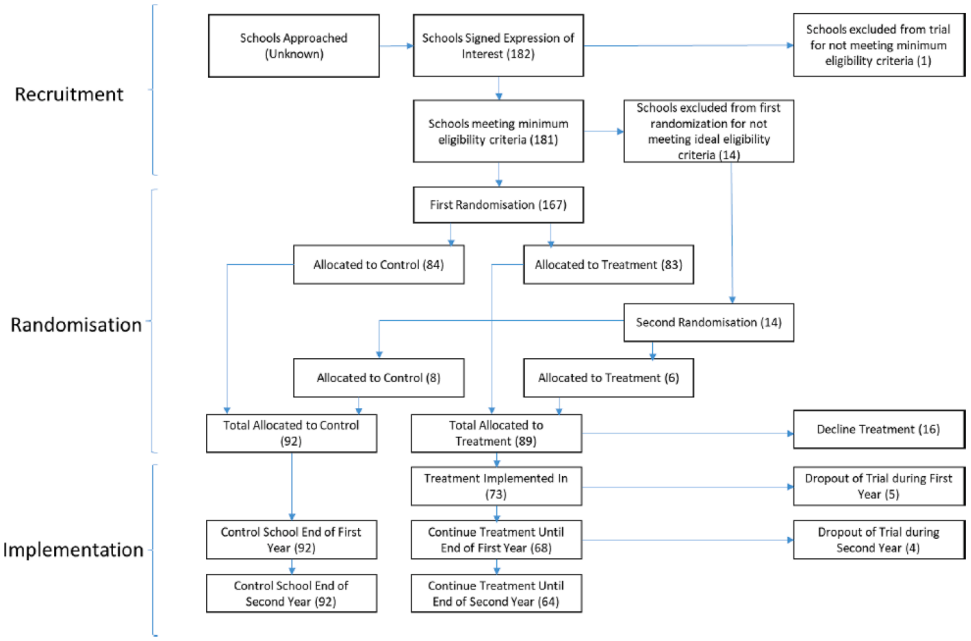

Fig. 3. Consort flow diagram

16 no longer wished to take part, leaving 67 treatment schools.[21]

The 14 schools that were excluded from the first round of randomisation were then randomized into treatment and control groups. Pairs were randomly generated within reason for initial exclusion. For schools in pair-franchises, they were randomised as a pair, so that both schools were allocated to the same treatment status (two were assigned to treatment and four to control). Ultimately this resulted in 92 schools being allocated to control status and 89 allocated to treatment status, of which 73 initally participated in the program. Figure 3 presents the consort diagram, which traces the sample from recruitment, ramdomization to participation in the trial. During the course of the two year program five schools dropped out during the first year and four during the second year.[22] This meant that 64 schools of the 89 schools assigned to treatment actually went through the full two-year intervention.

Our main analysis sample consists of these 181 schools (92 allocated to control and 89 to treatment), but we also provide a parallel set of results for the intended (one class per grade) schools and the unintended (more than one class per grade) schools. Given the pair-wise randomisation we need to include pair fixed effects in all specifications, therefore this parallel analysis can only be on the subset of the 104 schools where both pairs were small (one class per grade) schools, or where both pairs were large (more than one class per grade) schools.[23] The intended sample consists of 25 pair-IDs where both schools had one class per grade. We have 25 control schools and 24 treated schools in our intended sample. In contrast the unintended sample consists of 28 "pairs" where both schools had more than one class per grade. We have 27 control schools and 28 treated schools in our unintended sample.

Compliance: There are two types of non-compliance in this trial, school level attrition from the experiment, and student level non-compliance. Even though we use treatment assignment in our analysis and have outcome measures from non-compliers, it is still important to examine if dropout is non-random, potentially affecting any results or generalizations thereof. As explained in the previous sub-section there was a deviation from our analysis plan, where 91 schools with more than one class per grade were recruited into the trial. We present estimates for the full sample, the intended sample of small school pairs and the unintended sample of large school pairs.

In Appendix Table A.1, we examine non-compliance in the form of school attrition. The odd numbered columns present the mean characteristics of the remainer schools, and the even number columns report the differences between remainers and dropout schools, conditional on the pair fixed effects used for the randomization, for the first (Panel A) and second (Panel B) cohorts. There is relatively little difference between compliers and non-compliers in terms of the share of the students eligible for FSM and certified to have a disability, or in terms of their prior test scores. However, the schools that attrit from the sample are more likely to have higher proportions of students with ESL, or are classified as minority students. These differences are driven by the larger (unintended) schools.

In addition to schools being assigned to treatment and not being treated, students could also be assigned to treatment (by being enrolled in a treated school) but not treated. At the end of each school year we received class lists from treated schools regarding which students were treated, which included their unique student identifier. This allowed us to identify which students were taught by lesson study teachers. Individual-level treatment can differ from school-level treatment for two reasons. First, because they are in a class that is lead by a non-observed teacher. This occurs when there is more than one class per grade; the program only involves three teachers and therefore in a school with two classes per grade, one class over the two cohorts would be left untreated. The NPD data does not allow us to determine how many teachers are in a school year, but there is indicative evidence that this is the case - the proportion of a cohort being treated only falls below 50 percent in treated schools that participated in the study when the cohort size was above 34. Secondly, some students joined the school during the final year of primary school, meaning they take the age-11 tests with the treated cohort, but were not exposed to a program teacher since the treatment would have occurred before they joined. As some classes in treated schools are not treated, the students within a year group that

[21] Of the 16 schools not accepting treatment, 8 provided no reason, 5 reported staffing issues, one school change of school priorities, one due to school inspection, and one stating that that they only had 2 percent FSM and so should not be included

[22] Three of these schools this was due to teacher turnover, two due to having a new headteacher, two provided no reason, and two due to having to prioitise OfSted inspections

[23] School-pairs that differ in size will be in different sub-samples and so will not contribute to the estimation of the main parameter. They will be absorbed by the pair effect, which would now only reflect one school

Table 2
Randomisation tests: cohort 1 and cohort 2

	(1) Sample (all)	(2)	(3) Intended	(4)	(5) Unintended	(6)
	Treated	Difference	Treated	Difference	Treated	Difference
Panel A: Cohort 1						
Age-7 Test	15.614	0.166	15.620	0.254	15.930	0.318
	[3.539]	(0.114)	[3.709]	(0.265)	[3.563]	(0.174)
Free School Meals	0.237	0.012	0.232	-0.042	0.234	0.032
	[0.425]	(0.014)	[0.422]	(0.023)	[0.423]	(0.021)
Special Edu. Needs	0.139	0.008	0.170	0.001	0.116	0.010
	[0.346]	(0.010)	[0.376]	(0.027)	[0.324]	(0.010)
Gender: Male	0.502	-0.001	0.502	-0.006	0.499	-0.004
	[0.500]	(0.011)	[0.500]	(0.029)	[0.500]	(0.016)
Minority	0.240	0.037	0.071	-0.012	0.302	0.040
	[0.427]	(0.024)	[0.258]	(0.016)	[0.459]	(0.034)
ESL	0.210	0.054	0.056	-0.013	0.279	0.095
	[0.408]	(0.022)	[0.230]	(0.017)	[0.499]	(0.030)
School Size	46.896	-0.293	22.944	0.064	62.240	0.731
	[24.641]	(2.450)	[4.661]	(1.110)	[25.295]	(4.304)
Panel B: Cohort 2						
Age-7 Test	15.823	-0.150	16.037	0.236	16.006	-0.187
	[3.455]	(0.127)	[3.329]	(0.221)	[3.575]	(0.211)
Free School Meal	0.240	0.014	0.210	0.063	0.254	0.070
	[0.427]	(0.014)	[0.408]	(0.024)	[0.435]	(0.022)
Special Edu. Need	0.126	-0.010	0.102	-0.026	0.126	-0.017
	[0.332]	(0.011)	[0.302]	(0.018)	[0.332]	(0.019)
Gender: Male	0.504	-0.007	0.499	-0.012	0.493	-0.019
	[0.500]	(0.009)	[0.500]	(0.019)	[0.500]	(0.015)
Minority	0.243	0.038	0.066	-0.041	0.309	0.077
	[0.429]	(0.024)	[0.248]	(0.021)	[0.462]	(0.031)
ESL	0.217	0.049	0.059	-0.023	0.291	0.097
	[0.412]	(0.023)	[0.236]	(0.017)	[0.455]	(0.027)
School Size	47.422	-2.262	24.234	0.189	62.501	-4.992
	[23.257]	(3.789)	[4.576]	(1.110)	[23.433]	(8.026)
Pair FX		X		X		X

Notes: Panels A and B show balancing at the student level for cohorts 1 and 2. Number of obs.: Full sample: cohort 1 (cohort 2) 6,436 (6,298). Intended sample: cohort 1 (cohort 2) 1045 (1089) Unintended sample: cohort 1 (cohort 2) 2865 (2687). Panels C shows balancing at the postcode district level of the schools, based on all propety transactions betwen Q1 2010 and Q3 2013 (pre-treatment) and neighbourhood characteristics in % from the UK Census 2011. Standard deviations of variables shown in square parenthesis, standard errors clustered at the school level shown in round parenthesis.

receive treatment might be non-random, a possibility that we explore in Section 6.

4. Empirical approach

Prior to conducting the RCT we pre-committed to a set of specifications and outcome measures in a Statistical Analysis Plan (SAP)[24], which was written three months before the beginning of the trial. The purpose of the SAP is to minimize conscious or sub-conscious decisions being made on the basis of results seen. The SAP contains details of the study design, sample size, randomization, chosen outcome measures, methodology and analysis plan, subgroup analysis. We now follow exactly the evaluation strategy that we set out initially and indicate the few cases where we deviate.

Our primary analysis is conducted on an 'intention to treat' (ITT) basis. Specifically, we build up to from a univariate specification, only controlling for school assignment to treatment D_s, to the following model

$$Y_{ips} = \alpha + \beta D_s + X_{it}^{'}\delta + \pi_p + \varepsilon_{ips} \quad (1)$$

where the dependent variable Y_{ips} is pupil i's age-11 test score, in school pair p from school s. These students took their age-11 tests in the academic years 2014/15 (cohort 1) and 2015/16 (cohort 2). There are two systematic differences between these cohorts. First, students from the first cohort are only taught by teachers trained in the program for one year, whereas students from the second cohort are taught for two. Second, teachers will be more accustomed to the system by the second year therefore the second cohort are taught by teachers more experienced in the program. To account for these differences the model is estimated for each cohort separately.

β is our main parameter of interest and reflects the mean difference between those assigned to treatment and control groups for each cohort, where the treatment includes five days of teacher training over a two year period, along with three lessons being observed by two peers, and observing six lessons each year. With successful randomization, a direct comparison of the means should be sufficient for determining the effect size. To improve the efficiency of the estimations we include X_{is} a vector of pupil characteristics. These are the student's average age-7 test scores (across maths and reading), and indicators for gender, special educational needs, English as a Second Language (ESL), ethnic minority status (defined by the Department for Education as Non-White British) and Free School Meal (FSM) status. Given the pair-wise randomization structure, here we also include pair-fixed effects. Throughout the analysis all standard errors are clustered at the school level.[25]

As noted previously, some schools that were assigned to treatment dropped out of the program. We therefore also estimate LATEs via two-stage least squares (2SLS), where initial treatment allocation is used as an instrument for actual receipt of the intervention. It thereby scales the ITT estimate, by accounting for the non-compliance of some schools or

[24] This can be found at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/lesson-study/

[25] For the main results in Table A.3 we provide simulated Fisher exact p-values (see also Appendix Figures A.2 and A.3). Due to the large sample size of this trial, these are very similar.

students. The actual receipt of the intervention is defined in two ways. First, at the school cohort level (T_s), where we define a school to be treated if we received confirmation from the school at the end of each academic year that they participated. Second, at the student level (T_{is}), if we received confirmation from the school that the student was taught by an observed teacher.[26]

$$T_{is} = \alpha + \beta_1 D_s + X_i^{'}\delta_1 + \rho_p + \varepsilon_{ips} \quad (2)$$

$$Y_{ips} = \alpha + \beta_2 \widehat{T}_{is} + X_i^{'}\delta_2 + \pi_p + \tau_{ips} \quad (3)$$

We estimate the specfications 1 to 3 using the whole sample, and a parallel set of results using the intended sample (small schools) and unintended samples (large schools) as defined in section 3.5.

5. Results

5.1. Balance at baseline

Before presenting the effects of the progam on student outcomes, Table 2 shows summary baseline statistics for the treatment and control schools, for the whole sample, intended and unintended samples. We show both the average school characteristics for treated schools and the difference between the treatment and control for a wide range of student characteristics, for cohort 1 (Panel A) and cohort 2 (Panel B). There are no significant differences in these characteristics between the treatment and control groups in our full sample, indicating our randomisation generated balanced treatment and control groups. Looking at our intended (small schools) and unintended (large schools) samples, we do observe some small differences, with treated students more likely to be free school meal recipients in both the unintended and intended sample in cohort 2. As expected schools in the intended sample are considerably smaller than the unintended sample (24 versus 62), but within each type the treated and and control schools are not significantly different in grade size. Finally notice that there are no significant differences in student intake across treatment and control schools (age-7 test scores).

In addition to establishing if the student composition is similar in the treatment and control groups, we also show that the schools themselves are similar and are situated in similar neighbourhoods. We obtain characteristics about schools neighbourhoods from the 2011 Census (2 years before the program) at the postcode district level. There are 2269 postcode districts in England, with a median number of 9500 households. The data we take from the census includes the proportion of local population which; is non-white, has post-secondary qualification (level 4+), are homeowners, and are social-renters. In addition, to generate a broader and more continuous measurement of neighbourhood-level characteristics we generated a price index from UK Land Registry data. This is based on all property transactions between Q1/2010 to Q2/2013 (the quarter before the program started). The index is created at postcode district level by residualizing ln(price) by quarter*year dummies and a control for property type. In addition to neighbourhood characteristics, we also test to see if the treated schools are different in their type of school management e.g. Academy Converter, Community, Foundation, Voluntary Aided, Voluntary Controlled. These neighbourhood characteristics and differences are presented in Appendix Table A.2. We find that the treated and control schools are similar in terms of neighbourhood chracteristics and type for the main, intended and unintended samples.

[26] In our SAP we also specified an alternate specification, in which we exploit the panel nature of the administrative data, which increases our sample size dramatically, and allows us to perform a difference-in-differences analysis. We do not present these results here since gains in precision from this exercise are negligible.

Table 3
Main results: cohort 1 and cohort 2

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample (all)		Intended		Unintended	
Panel A: Cohort 1						
Test Score	1.301	0.270	-1.806	-3.082	3.670	1.890
	(1.089)	(0.990)	(2.005)	(1.548)	(1.881)	(1.773)
Maths	2.064	0.897	-2.068	-3.414	5.019	2.951
	(1.240)	(1.130)	(2.417)	(1.869)	(2.128)	(2.024)
Reading	0.538	-0.357	-1.544	-2.750	2.321	0.828
	(1.054)	(0.962)	(1.924)	(1.688)	(1.811)	(1.676)
SPAG	1.229	-0.237	-4.207	-5.534	4.172	1.669
	(1.222)	(1.115)	(2.122)	(2.059)	(2.024)	(1.853)
Panel B: Cohort 2						
Test Score	-0.004	0.597	-4.124	-5.365	3.919	4.897
	(1.226)	(1.132)	(2.109)	(1.635)	(2.028)	(1.875)
Maths	0.493	1.003	-4.258	-5.497	5.541	6.253
	(1.413)	(1.299)	(2.551)	(2.090)	(2.302)	(2.219)
Reading	-0.502	0.192	-3.991	-5.234	2.296	3.541
	(1.162)	(1.099)	(1.932)	(1.487)	(1.921)	(1.675)
SPAG	-0.925	-0.585	-6.237	-7.286	2.452	2.674
	(1.133)	(1.105)	(2.632)	(2.422)	(1.462)	(1.678)
Pair FX	X	X	X	X	X	X
Student controls		X		X		X

Notes: This tables shows results of the intervention at age-11 on average english and maths test scores [Test Score (age-11)], maths test scores, reading test scores, scores for spelling and punctuation and grammar [SPAG], separately for cohort 1 [Panel A] and cohort 2 [Panel B]. Standard errors clustered at the school level shown in round parenthesis.

5.2. Effects on pupil attainment: main results

The main results are presented in Table 3, where we report ITT estimates on national test percentile rank. We present results with pairwise fixed effects (columns 1,3,5) and with the addition of student controls (columns 2,4,6) for our full, intended and unintended samples. We estimate effects of the intervention on a combined test score measure, maths test scores, reading test scores and a score for spelling, punctuation and grammar. All scores are percentalized at the cohort-by-subject level so that these measures have an average of about 50 and standard deviation of about 28.8.

As Table 3 shows, we observe no significant effects for any of the outcomes and across both cohorts in the full sample of schools. This is shown in columns 1 and 2. Notably, these estimates - although never statistically different - are a little sensitive to the inclusion of contorls for cohort 1. This is a reflection of the small imbalances discussed above. Conditional on pair fixed effects and student covariates, students in schools assigned to treatment scored on average 0.27 percentage points higher on centralised age-11 exams in the first cohort, and 0.6 percentage points higher in the second cohort, neither effect is statistically significant. We can reject that the average impact is larger than 2.2 and 2.8 percentage points for cohorts one and two respectively.

Columns 3 and 4 show results for the intended sample of schools, by estimating ITT effects for schools where both schools in each randomisation pair have only one class per grade. In contrast to the full sample, we find clear evidence of negative effects of the intervention. For cohort 1, students in schools allocated to the treatment score on average 1.8 percentiles lower than the control schools, although not statistically significant having as Standard Error (SE) of 2.00. The effect increases and becomes more precise with the inclusion of student controls - mainly stemming from the prior student age-7 test scores. Students in schools allocated to treatment score 3.1 percentiles worse than those in control schools (SE 1.5). This is equivalent to 10 percent of a standard deviation. This effect is seen over all tested subjects - math, reading and SPAG - although only statistically significant at traditional levels for the latter. For the second cohort, who were taught by Lesson Study teachers for

Table 4
IV analysis

	(1) School LATE	(2)	(3)	(4) Student LATE	(5)	(6)
	Sample (all)	Intended	Unintended	Sample (all)	Intended	Unintended
Panel A: Cohort 1						
Test Score	0.335	-3.969	2.384	0.405	-4.160	2.978
	(1.226)	(1.958)	(2.164)	(1.482)	(2.061)	(2.754)
Maths	1.114	-4.397	3.724	1.345	-4.608	4.652
	(1.396)	(2.373)	(2.483)	(1.698)	(2.497)	(3.197)
Reading	-0.444	-3.542	1.044	-0.536	-3.712	1.305
	(1.203)	(2.121)	(2.058)	(1.449)	(2.276)	(2.585)
SPAG	-0.294	-7.127	2.106	-0.355	-7.470	2.631
	(1.388)	(2.660)	(2.283)	(1.673)	(2.781)	(2.899)
First Stage	0.805	0.776	0.793	0.667	0.741	0.634
	(0.033)	(0.065)	(0.075)	(0.032)	(0.062)	(0.100)
Panel B: Cohort 2						
Test Score	0.747	-6.880	6.377	0.886	-7.590	8.357
	(1.406)	(2.235)	(2.293)	(1.679)	(2.497)	(3.266)
Maths	1.253	-7.050	8.144	1.487	-7.776	10.672
	(1.608)	(2.813)	(2.673)	(1.928)	(3.106)	(3.844)
Reading	0.240	-6.711	4.611	0.285	-7.403	6.042
	(1.372)	(2.004)	(2.097)	(1.630)	(2.276)	(2.915)
SPAG	-0.731	-9.344	3.483	-0.868	-10.307	4.564
	(1.386)	(3.199)	(2.118)	(1.641)	(3.614)	(2.829)
First Stage	0.800	0.780	0.768	0.674	0.707	0.586
	(0.035)	(0.063)	(0.063)	(0.033)	(0.066)	(0.056)

Notes: Columns (1) to (3) show estimates of the causal effect of the teacher training program where assignment to treatment is used to instrument for school-level take-up. Columns (4) to (6) show results when random assignment to the treatment is used as instrument for actual student-level take-up. Pair-FX, Age-7 test scores and student demographics are included as controls. Number of observations: Full sample: cohort 1 (cohort 2) 6,436 (6,298). Intended sample: cohort 1 (cohort 2) 1045 (1089) Unintended sample: cohort 1 (cohort 2) 2865 (2687). Standard errors clustered at the school level in parenthesis.

twice as long, the effects of the intervention are larger. The average maths and reading performance in these schools is -5.37 (SE 1.64) percentiles lower. Moreover, the impacts are statistically significant for each subject, with maths and reading performance being equally effected (-5.5 and -5.23).

The final two columns (5 and 6) show results for the unintended sample of large schools. Here, we find positive effects throughout. Conditional on student characteristics, student in large schools who were assigned to treatment score 1.8 (SE 1.77) percentiles higher in the first cohort, and 4.9 (SE 1.90) percentiles higher for the second cohort. The statistically significant results of the second cohort are equivalent to a 17 percent of a standard deviation increase in student performance. As with the unintended sample, these effects are not significant for individual subjects for the first cohort. For the second cohort there are significant positive impacts on reading, and math with larger (but not significantly different) impact on maths performance.

Teacher observation is more effective in larger schools, and the differences get larger over time. It is also of note that the results for small schools are negative rather than just null (and that this negative increases over time). In Section 6 we discuss a number of potential mechanisms for what could be driving these differences by school size.[27]

5.3. IV analysis

The estimates presented so far are intention to treat effects and so will underestimate the impact of those that actually experienced the program: some students in the treated group were not treated, either because the entire school dropped out of the experiment, or because they are in a larger school where not all the classes in a grade were treated. We now present 2SLS estimates by instrumenting actual participation status with random school assignment. Columns 1 to 3 of Table 4 defines participation at the school-level, Columns 4 to 6 instead defines participation at the student-level. This student level treatment information comes from reports from the treated schools each year (see Section 6 for details).

The purpose of presenting two sets of 2SLS estimates is because some students are not treated in treated schools, and therefore the role of spillover effects should be considered. There are many potential mechanisms for these spillovers, which have the potential for negative or positive externalities. The potential spillovers across classes is the reason why we include the school and student level definitions of treatment side-by-side. If one believes spillovers are important then the correct definition of treatment is at the school level, and one should focus only on the estimates in columns 1 through 3, which reflects the average improvement for all students in schools who were selected and participated in the treatment. In contrast, if one was to believe there were no spillovers across classes, then the focus should be on the estimates in columns 4 through 6. Here the exclusion restriction is that gains are assumed to be concentrated in only the treated classes. The student level estimates are important as arguably they are closer to the improvement that would occur if this training was to be implemented in all classes. Again, all results are reported for the overall sample, the intended sample of single class per cohort schools and the unintended sample of larger schools.

Before discussing the magnitude second stage estimates, we focus on first stage estimates which provide us with new information. First, the first stage parameters for the school level instrument are larger for the school level treatment status than the student level treatment status. This reflects the fact that not all students in a treated school participate in the program. Second, the differences between the first stages using school assignment between intended and unintended samples are small, establishing there is no differential attrition between small and large schools. Third, for the student analysis the first stage is smaller for the unintended sample, compared to the intended sample. This is because in large schools there are multiple classes per grade and so not all of the students are in a treated class.

Given the size of the first stages the second stage estimates are as expected i.e. the same reduced form effect is divided through by the differently-sized first stages. The negative estimates for the intended

[27] In addition to these ITT-results, following our RCT protocol, we include cross-sectional overall effects in the Appendix Table A.3, with inference based on Fisher-exact p-values.

Table 5
Heterogeneity - School Level

	(1) Test Scores	(2)	(3) Maths	(4)	(5) Reading	(6)	(7) SPAG	(8)
	High/Yes	Low/No	High/Yes	Low/No	High/Yes	Low/No	High/Yes	Low/No
Panel A: Cohort 1								
Quality of School Leadership	0.031	-0.100	0.015	-0.144	-0.072	-0.036	-0.085	0.046
	(0.069)	(0.071)	(0.066)	(0.079)	(0.065)	(0.065)	(0.059)	(0.095)
Teaching Quality	-0.058	-0.102	-0.034	-0.145	-0.122	-0.039	-0.126	0.043
	(0.071)	(0.071)	(0.057)	(0.079)	(0.055)	(0.065)	(0.053)	(0.095)
Safety and Behaviour	0.164	0.012	0.203	0.018	0.085	0.004	0.221	0.011
	(0.122)	(0.051)	(0.120)	(0.049)	(0.132)	(0.050)	(0.151)	(0.056)
Pupil Attainment	-0.083	-0.135	-0.048	-0.189	-0.103	-0.056	-0.151	-0.136
	(0.064)	(0.070)	(0.062)	(0.077)	(0.061)	(0.068)	(0.055)	(0.074)
Panel B: Cohort 2								
Quality of School Leadership	-0.011	0.073	-0.009	0.097	-0.011	0.035	-0.093	0.203
	(0.078)	(0.077)	(0.075)	(0.079)	(0.072)	(0.072)	(0.062)	(0.122)
Teaching Quality	-0.033	0.073	-0.019	0.098	-0.040	0.034	-0.086	0.201
	(0.065)	(0.077)	(0.063)	(0.079)	(0.062)	(0.072)	(0.051)	(0.122)
Safety and Behaviour	0.140	0.033	0.238	0.001	0.022	0.061	0.083	0.007
	(0.109)	(0.063)	(0.118)	(0.062)	(0.121)	(0.056)	(0.144)	(0.052)
Pupil Attainment	-0.048	-0.004	-0.033	0.008	-0.053	-0.015	-0.077	0.085
	(0.070)	(0.074)	(0.069)	(0.071)	(0.063)	(0.078)	(0.057)	(0.115)

Notes: This tables shows standardised estimates for subsamples of randomisation pairs classified according to the indicators in columns 1. All specfications include pair FX and age-7 test scores as controls. Standard errors clustered at school level in parenthesis.

Table 6
Heterogeneity - Individual Level

	(1) Test Scores	(2)	(3) Maths	(4)	(5) Reading	(6)	(7) SPAG	(8)
	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction
Panel A: Cohort 1								
FSM	0.023	-0.015	0.046	-0.007	-0.007	-0.016	0.031	-0.064
	(0.041)	(0.043)	(0.043)	(0.045)	(0.038)	(0.045)	(0.044)	(0.048)
ESL	0.011	-0.014	0.025	0.035	-0.005	-0.063	-0.031	0.135
	(0.043)	(0.067)	(0.044)	(0.069)	(0.038)	(0.068)	(0.044)	(0.080)
Minority	0.015	0.004	0.026	0.062	-0.000	-0.056	-0.032	0.164
	(0.042)	(0.095)	(0.043)	(0.063)	(0.037)	(0.061)	(0.042)	(0.071)
Low age-7 test score	0.013	0.038	0.039	0.032	-0.019	0.049	-0.007	0.116
	(0.043)	(0.055)	(0.044)	(0.058)	(0.039)	(0.054)	(0.046)	(0.057)
Gender: male	0.001	0.033	0.031	0.023	-0.033	0.042	-0.012	0.051
	(0.043)	(0.031)	(0.047)	(0.039)	(0.040)	(0.035)	(0.045)	(0.034)
Panel B: Cohort 2								
Share Free School Meals	0.035	-0.013	0.051	-0.004	0.017	-0.018	0.003	-0.014
	(0.048)	(0.045)	(0.050)	(0.046)	(0.042)	(0.048)	(0.044)	(0.047)
ESL	0.012	0.059	0.018	0.104	0.009	0.004	-0.034	0.098
	(0.048)	(0.074)	(0.049)	(0.080)	(0.042)	(0.069)	(0.043)	(0.087)
Minority	0.017	0.058	0.030	0.077	0.005	0.033	-0.025	0.069
	(0.051)	(0.067)	(0.052)	(0.068)	(0.045)	(0.069)	(0.045)	(0.070)
Low age-7 test score	0.032	-0.011	0.044	0.036	0.015	-0.022	-0.007	0.032
	(0.048)	(0.055)	(0.051)	(0.064)	(0.041)	(0.052)	(0.045)	(0.059)
Gender: male	0.063	-0.062	0.095	-0.088	0.021	-0.019	0.013	-0.031
	(0.046)	(0.036)	(0.049)	(0.040)	(0.043)	(0.039)	(0.042)	(0.035)

Notes: This tables shows estimates for main effects and interactions for age-11 outcomes in overall test scores (col 1-2) maths (col 3-4), reading (col 5-6) and spelling, punctuation and grammar (col 7-8). Columns 3-8 is not part of the pre-registered analysis plan of the intervention. All specfications include pair FX and age-7 test scores as controls. Estimates are standardized. Standard errors clustered at school level in parenthesis.

sample and the positive estimates for the unintended samples are reinforced, and estimates are larger for the second cohort who went through treatment for two years, instead of one. Instrumenting for actual school participation with assignment, we find that schools in the intended sample score 4 percentile points lower in the first cohort, and 6.8 percentile points lower in the second cohort. This is equivalent to scoring 13 and 23 percent of a standard deviation lower respectively. In contrast large schools gained by 8 percent and 22 percent of a standard deviation in cohorts one and two respectively. Note that the effects on maths scores are larger, in particular for the second cohort, where students in treated (large) schools outperform students in untreated (large) schools by thirty percent of a standard deviation in the national externally marked test score. Given the negative estimates for the intended sample, and the large positive effects for the unintended sample, it is clear that teacher training can have significant effects on student learning - but the sign of the effect depends on the setting.

6. Mechanisms

We now explore four potential mechanisms that may explain these differing effects between small and large schools; 'compositional', 'selection', 'matching' and 'disruption'.

First considering composition, it may simply be the case that larger schools are systematically different in other dimensions that may be responsible for the larger effect sizes in such schools. For example, larger schools may differ in terms of their quality of school leadership, teaching quality, safety and behavior. In Table 5 we explore these school level characteristics, using information from government (Ofsted) inspections of

the schools from before the intervention. We categorise the school Ofsted-ratings into high (Outstanding, Good) or low (Satisfactory, Inadequate), and estimate ITT effects on subsamples of high and low rated schools. Note, as we include randomization-pair fixed-effects in all specifications, only school pairs which have the same rating are used in these sub-sample analysis.[28] For neither the first nor the second cohort does splitting the sample by any of these four characteristics generate significant effects on average test scores - unlike school size. In addition to Ofsted ratings, we also categorise schools by student attainment (pre-treatment) in Key-Stage 2 performance by above or below the sample median (again categorised as high or low). Again, we find that the program is no more effective in previously high or low performing schools.

Exploring the heterogeneity of the effect using subsamples in this way makes few functional form assumptions, but lowers the power to reject that the effects are different. Therefore, in Panel A of Appendix Table A.4, we present results from an alternate approach which allows us to use the full sample using a numeric value for the school-level characteristic (Ofsted-ratings: Outstanding-4, Good-3 Satisfactory-2 Inadequate-1; School performance: Standardized student achievement) and interacting it with treatment status. Again, we find no evidence of heterogeneity for any characteristic other than school size.

In Table 6, we investigate the possibility that the heterogenous results are driven by the composition of the student body by exploring student-level heterogeneity. If the program is more effective with certain types of students, then the positive effect in large schools could be driven by schools having a higher proportion of this type of student. The parameters in Table 6 are of the school level treatment, plus its interaction with five binary student characteristics (FSM, ESL, ethnic-minority, male, and low prior achievement). The vast majority of the interaction terms are insignificant, implying that school composition does not play a large role in the divergent results. Out of the forty interaction terms, two are statistically significant at the five percent level (SPAG for minority students in the first cohort, and male for math in the second cohort). Given the number of coefficients tested we would expect this number of coefficients to be significant at the 5 percent level, and so we conclude that there is little evidence for student level heterogeneity. Panel B of Appendix Table A.4 presents treatment status interacted with the grade share of different student characteristics. Consistent with Table 6, we find no evidence of heterogeneity by student characteristics.

Second, it might be the case that selection effects are driving the result we see - having more than one class per grade would allow larger schools to select the teachers and students that participated in the program, and so may select those with the highest marginal benefit. To explore this, we use the class lists of students that were taught by trained teachers to determine if the excluded classes are systematically different to the treated classes in large schools. Appendix Table A.5 presents the characteristics of treated and non-treated students within treated schools. There are almost no differences between treated and untreated students, including on student prior attainment, implying that classes were not chosen. Critically there are no differences in prior attainment of pupils in treated and un-treated classes within treated schools. We take this as strong evidence that schools were not actively sorting students into treated classes. Of course, it is possible that there are unobservable factors behind which classes were chosen. However, the fact that our sample is balanced on several characteristics and especially pupil prior test scores is very reassuring. Our administrative data does not include information on teachers, and so we are unable to test if trained teachers are systematically different from trained teachers in treated schools.

A third potential mechanism is that the gains in larger schools could be driven by 'matching' effects. Here, we assume that for peer-to-peer feedback programs to work there must be sufficient heterogeneity of teacher quality among the participants so that teachers can learn from each other. This situation is more likely in larger schools with more teachers, than smaller ones. This hypothesis is difficult to test, since we do not have data on teachers themselves. However, it is backed up by the literature, e.g. Papay et al. (2018) find positive co-worker effects among pairs of teachers who were purposefully paired up based on previous effectiveness measures. The results described above, in Appendix Table A.4, also provide suggestive evidence for this hypothesis - when including interactions of treatment status with cohort size in our specification, we find that the treatment effect increases linearly with the size of the school cohort. The negative treatment effect becomes positive with cohort size, with a tipping point of around 42 students, which is approximately one and a half classes. This result implies that the more teachers there are in a school, the more successful the treatment effect is.

Finally, our results are consistent with the intervention being more disruptive in smaller schools than in larger schools. This could be both from the initial introduction and the perpetual implementation of the program. As previously mentioned, the intervention involved 5 days of training. The disruption caused by taking teachers out of class to attend the training could have been more severe in smaller schools than larger schools. Moreover, the accompanying process evaluation of the program found evidence indicating that small schools faced more organizational challenges in implementing the program. A key finding of the process evaluation was that "Teachers in the case study schools felt that the Lesson Study cycle was best delivered in a concentrated period of time, so that planning, delivery and reflection could be best co-ordinated and carried out effectively. This presented some organizational challenges, particularly in smaller schools." A possible reason why smaller schools had trouble coordinating the timings of the observations could be due to a higher proportion of the school teaching body having to be in the same classroom at the same time.

That the negative effect of the treatment for small schools increases over time implies that in addition to any initial disruption, smaller schools experienced more disruption in the second year of the program. This implies that the program is less effective than business as usual, in each year of implementation. In addition to any disruption effects, in schools with one class per grade, teachers would not have had the option to opt in to the program. This in itself may demotivate teachers and perpetually diminish the impact of the program.

In summary, we find little evidence that the large differences in the effectiveness by school size are primarily driven by other school characteristics, student selection or one-time disruption effects. On the other hand, we find that a continuous measure of school size provides the same type of heterogenous results. This is consistent with matching effects (sufficient variation in teacher quality within groups), and disruption effects. For example, in small schools the teachers were particularly poorly matched, and actually gave each other harmful advice. Or, that teacher training is typically better organised and productive in smaller schools, and the relatively ineffective intervention took teachers away from this. Regardless of whether it is due to matching or disruption, the finding that the negative effects increase over time implies that the peer-to-peer training is less effective than business as usual in smaller schools.

7. Discussion and Conclusions

Teacher peer observation is a popular practice, adopted by schools either as a means to identify productive teachers, or to improve their existing labor force. By implementing a large-scale randomized control trial, with high fidelity, across primary schools in England, we attempt to provide robust evidence on the efficacy of teacher peer observation as a teacher development tool.

Our main finding is that on average there is no significant effect of peer-to-peer training on student achievement in English primary schools. We can reject that the average effect of Lesson Study is larger than 0.077σ for the first cohort and 0.098σ for the second cohort. This is

[28] The pair fixed effect includes one treated and one control school. Having only one in a sample will absorb any treatment effects.

considerably smaller than the average effect of reducing class size from 22 to 15 found in the Project STAR experiment (0.15-0.2σ Krueger & Whitmore (2001)), and of the average effect of teacher observation conducted by expert external evaluators which potentially had consequences (0.11σ (Taylor & Tyler, 2012)).

However the average effect masks a signficant heterogeneity by school size. In smaller schools where no teacher selection was possible since there was one class per grade, the program had a clear negative impact on student test scores. Here we can reject positive effects for both cohorts, and by the second cohort the average impact was -0.186σ allowing us to reject effects larger than -0.075σ. By contrast, in larger schools, there were gains of 0.07σ for the first cohort and 0.17σ for the second.

There are four potential reasons for this; 'compositional', 'selection', 'matching' and 'disruption' effects. We provide indicative evidence against the first two explanations, and have qualitative evidence in favour of the fourth. While cannot say with certainty which of these is playing the dominant role, we can say that a "one size fits all" approach to teacher peer observation, with teachers instructed to observe each other regardless of the setting, is ineffective, and may even lead to negative impacts. Moreover we find that a key determinant of effectivness is school size. If the matching effect is the dominant driver of the heterogeneity by school size, with limits on the variation in teacher quality within a training group in smaller schools, then one potential solution would be to put outside experts into the groups, to alleviate this constraint. However, this would then no longer be considered to be peer-to-peer learning.

The positive effect we observe in larger schools is in line with findings from Taylor and Tyler (2012), whose evaluation of a one year intervention with three observations finds an impact of 0.112σ in maths achievement in the first year after the observations, and effects of 0.158σ two years after the observations.[29] However, for our full sample we can reject effects of up to 11.04 for cohort 1 and 12.62 percent of a standard deviation for cohort 2.[30] Unlike our setting, their program featured external observers (hence teacher learning was more likely) and took place in middle schools which are larger than English primary schools.

In contrast to other forms of school policies to improve student outcomes, peer-to-peer teacher training in large schools comes out comparatively well. We can reject that the average effect of lesson study in large schools is smaller than the average effect of; removing two disruptive students from an elementary class of 25 (Carrell, Hoekstra, & Kuka, 2018), the provision of home computers (Fairlie & Robinson, 2013); or from having a one-standard deviation better teacher peer present in the school (Jackson & Bruegmann, 2009).

Our positive finding in large schools is also in line with recent evidence from Burgess et al. (2021) who evaluate a very similar program to ours, in UK high schools. High schools are typically significantly larger than the primary schools of our setting, so this lends further weight to our hypothesis that the increased probability of high quality observer teachers increases the likelihood of a successful outcome. As such, this paper brings new evidence on the potential mechanisms through which the Cincinnati and Burgess studies may have generated positive results, using a large sample RCT.

Our study does have two limitations which should be noted.

First, a pre-condition for being able to recruit so many schools in the English context were assurances that we would not use teacher-level data. As a result we cannot examine teacher-level improvements directly. In practice, baseline measures of individual teacher-level effectiveness are often not available to education policy makers, who are often making decisions about the implementation of teacher training programs. Moreover, even if such information were available, its usefulness would be limited in small schools, where there is little choice for teacher selection.

Second, many schools already implement some form of peer-to-peer feedback, albeit in a less structured and comprehensive way. Teacher interactions and co-worker learning is likely taking place even in the absence of the Lesson Study intervention, e.g. Jackson and Bruegmann (2009). Our research cannot quantify what the impact of the intervention would be compared to schools who do not carry out any of these activities, rather it is a comparison to business as usual.

Our results provide a robust evaluation of Lesson Study, a peer-to-peer observation program used in over fifty countries. We can reject that the program has meaningfully large effects on student achievement in primary schools in England who have above average shares of FSM students. Given our research design, these results are generalizable to the implementation of Lesson Study more generally as the sample consisted of schools who wanted to participate in the program. Moreover, in larger schools, where leadership had the option of the selection of teachers and students, this also replicates the practical realities of implementation. The use of teacher observation and feedback is gaining traction and there are many commonalities in approaches used across schools in the UK and internationally. We believe that the results of this research are highly relevant for schools carrying out these activities. The combination of the pair-wise RCT design and access to administrative student records and assessments makes this study compelling. As described above, our results indicate that teacher observation and feedback is not effective in every setting. We therefore conclude that policy makers need to pay close attention to heterogeneity in effects of educational interventions. This cautions against the notion that the same policy intervention can generate identical effects across different teachers and schools, and centralised one-size-fits all interventions in education policy related to teacher training. We believe exploring these heterogeneities and the scalability of positive effects found in some settings is an important avenue for future research.

CRediT authorship contribution statement

Richard Murphy: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Felix Weinhardt:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Funding acquisition. **Gill Wyness:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Funding acquisition.

Appendix A. Participating Local Authorities

LA Code	Name
341	Liverpool
342	St Helens
343	Sefton
344	Wirral

(*continued on next page*)

[29] This is best compared to our estimates from Table 3 Column 6 for maths outcomes with effects of 10.2 percent of a standard deviation for cohort 1, and 21 percent of a standard deviation for cohort 2.

[30] This is best compared to our estimates from Table 3 column 2 for maths outcomes.

(*continued*)

352	Manchester
353	Oldham
354	Rochdale
356	Stockport
357	Tameside
821	Luton
823	Central Bedfordshire
867	Bracknell Forest
873	Cambridgeshire
874	Peterborough, City of
878	Devon
879	Plymouth, City of
888	Lancashire
896	Cheshire West and Chester

Figures and Tables

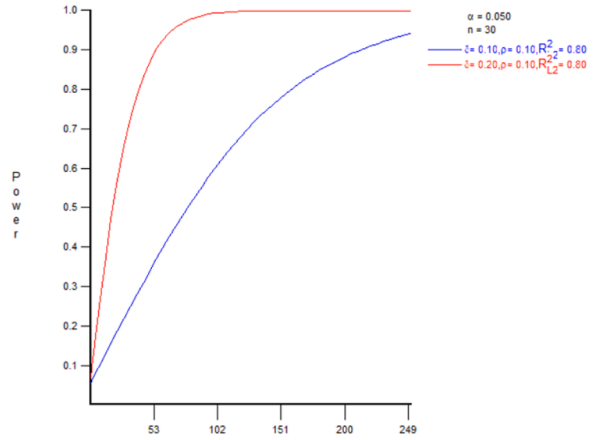

Fig. A.1. Power calculations, pre-trial Notes: These power calculations assumes; 30 students per class; an intra class correlation of 0.1; 0.8 of variation explained by student observables and randomization-pair effect. The blue line indicates power with effect size of 0.2σ, the red line effect size of 0.1σ.

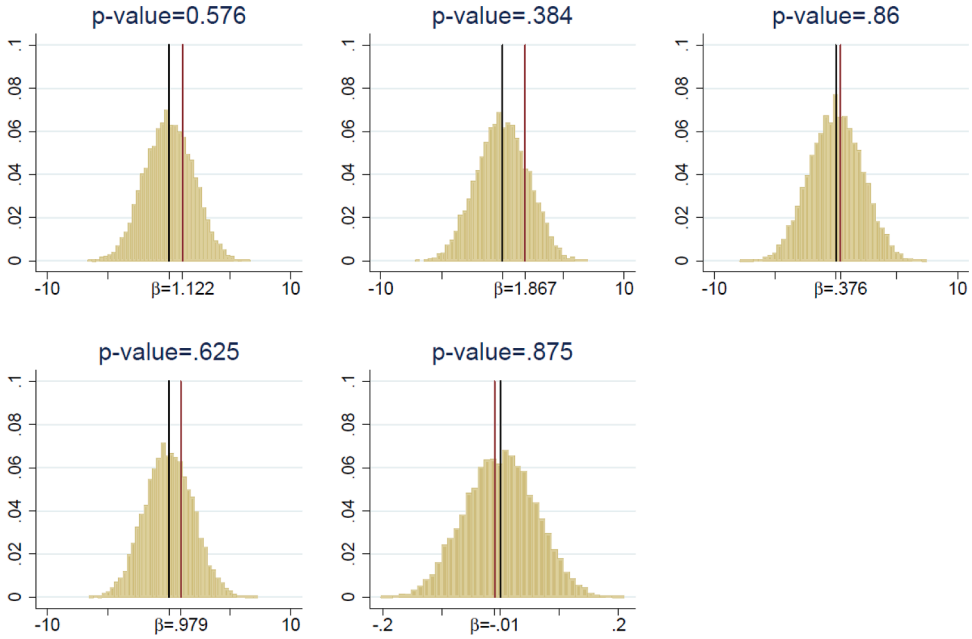

Fig. A.2. Simulated fisher exact p-values, cohort 1 Notes: To obtain these distributions, treatment status was randomly assigned within school pairs. 10,000 simulations each. This is for Table 3, Panel A.

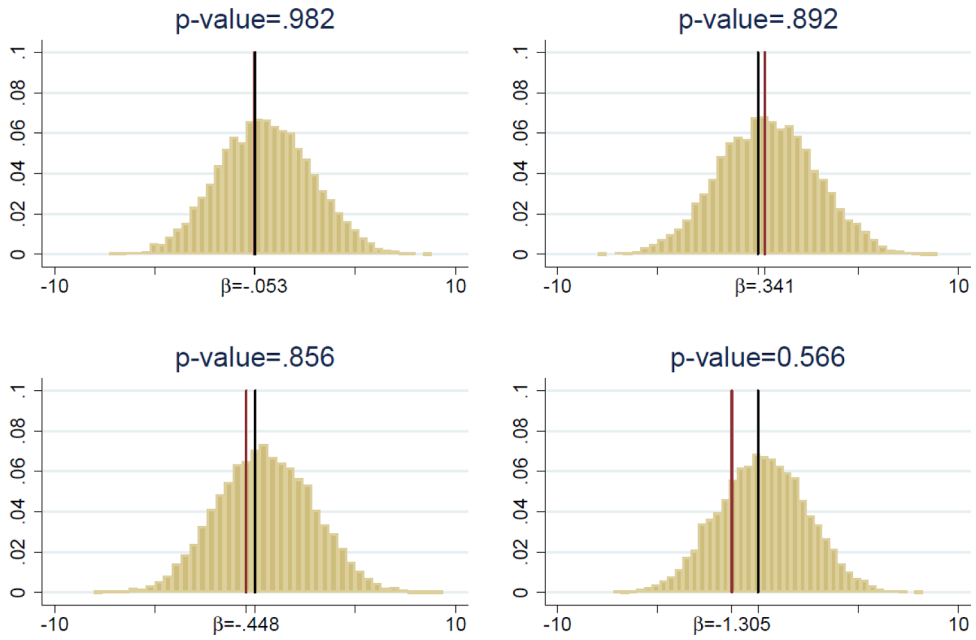

Fig. A.3. Simulated fisher exact p-values, cohort 2 Notes: To obtain these distributions, treatment status was randomly assigned within school pairs. 10,000 simulations each. This is for Table 3, Panel B.

Table A.1
Analysis of School-Level Dropout

	(1) Sample (all)	(2)	(3) Intended	(4)	(5) Unintended	(6)
	Stayer	Difference	Stayer	Difference	Stayer	Difference
Panel A: Cohort 1						
Age-7 Test	15.553	0.560	15.432	0.491	15.835	1.310
	[3.596]	(0.238)	[3.726]	(0.557)	[3.539]	(0.251)
Share Free School Meals	0.233	-0.002	0.280	-0.117	0.203	0.050
	[0.423]	(0.030)	[0.449]	(0.040)	[0.402]	(0.048)
Gender: Male	0.502	0.019	0.513	-0.021	0.501	0.009
	[0.500]	(0.026)	[0.500]	(0.055)	[0.500]	(0.039)
Share Special Edu. Needs	0.137	0.004	0.177	-0.012	0.110	0.023
	[0.344]	(0.018)	[0.382]	(0.041)	[0.313]	(0.028)
ESL	0.149	0.188	0.056	0.001	0.174	0.282
	[0.356]	(0.054)	[0.230]	(0.035)	[0.379]	(0.077)
Minority	0.185	0.204	0.081	0.030	0.221	0.252
	[0.389]	(0.055)	[0.273]	(0.024)	[0.415]	(0.076)
School Size	47.059	10.951	23.567	3.924	61.789	16.222
	[26.453]	(4.371)	[4.998]	(2.064)	[28.913]	(5.126)
Panel B: Cohort 2						
Age-7 Test	15.988	0.039	15.847	1.420	16.106	-0.094
	[3.596]	(0.292)	[3.398]	(0.376)	[3.456]	(0.512)
Share Free School Meals	0.231	0.023	0.237	-0.069	0.212	0.093
	[0.422]	(0.028)	[0.425]	(0.059)	[0.409]	(0.032)
Gender: Male	0.511	-0.057	0.506	-0.089	0.509	-0.060
	[0.500]	(0.024)	[0.500]	(0.032)	[0.500]	(0.035)
Share Special Edu. Needs	0.134	0.024	0.117	0.036	0.137	-0.011
	[0.341]	(0.023)	[0.332]	(0.033)	[0.343]	(0.049)
ESL	0.155	0.129	0.055	-0.068	0.178	0.242
	[0.362]	(0.057)	[0.229]	(0.020)	[0.382]	(0.078)
Minority	0.186	0.152	0.080	-0.070	0.436	0.248
	[0.389]	(0.058)	[0.272]	(0.012)	[0.142]	(0.073)
School Size	50.151	9.332	24.345	1.234	68.749	13.113
	[33.073]	(4.782)	[4.510]	(2.251)	[23.433]	(5.817)
Pair FX		X		X		X

Notes: Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

Table A.2
Randomisation tests at the school-level

	(1) Sample (all)	(2)	(3) Intended	(4)	(5) Unintended	(6)
	Treated	Difference	Treated	Difference	Treated	Difference
Panel A: Neighbourhoods of schools						
Property Price Index	12.109	0.003	12.128	0.014	12.101	-0.006
	[0.088]	(0.009)	[0.100]	(0.022)	[0.082]	(0.015)
Urban	0.843	0.047	0.760	0.091	0.889	0.074
	[0.366]	(0.044)	[0.010]	(0.092)	[0.320]	(0.075)
Level 4+ Qualification	0.053	0.003	0.048	0.005	0.066	0.005
	[0.024]	(0.002)	[0.021]	(0.005)	[0.033]	(0.005)
Ownership	0.627	-0.006	0.628	-0.005	0.606	-0.033
	[0.123]	(0.014)	[0.145]	(0.030)	[0.113]	(0.021)
Social Renter	0.194	0.012	0.196	0.025	0.198	0.022
	[0.100]	(0.010)	[0.117]	(0.018)	[0.087]	(0.013)
Panel B: School types						
Academy	0.034	0.000	0.000	0.000	0.111	0.000
	[0.181]	(0.024)	[0.000]	(0.000)	[0.320]	(0.076)
Community	0.528	-0.047	0.240	-0.318	0.593	-0.000
	[0.502]	(0.067)	[0.436]	(0.143)	[0.501]	(0.093)
Foundation	0.079	-0.012	0.120	0.000	0.074	0.000
	[0.271]	(0.031)	[0.332]	(0.066)	[0.267]	(0.054)
Voluntary Aided	0.258	0.023	0.440	0.182	0.185	0.074
	[0.440]	(0.060)	[0.507]	(0.143)	[0.396]	(0.075)
Voluntary Converter	0.101	0.035	0.200	0.136	0.037	-0.074
	[0.303]	(0.043)	[0.408]	(0.106)	[0.192]	(0.075)
Pair FX		X		X		X

Notes: Panel A shows balancing at the postcode district level of the schools, based on all propety transactions betwen Q1 2010 and Q3 2013 (pre-treatment) and neighbourhood characteristics in % from the UK Census 2011. Panel B shows school-level balancing by type. Differences are within paris, as by our randomization design. Standard deviations of variables shown in square parenthesis, standard errors clustered at the school level shown in round parenthesis.

Table A.3
Cross-sectional results

	(1) Treatment	(2) Control	(3) Difference	(4) Standardised	(5) Fisher p-value
Panel A: Cohort 1					
Test Score	47.25	46.13	1.12	0.044	
	(0.46)	(0.44)	(1.734)	(0.068)	0.376
Maths	48.13	46.26	1.87	0.066	
	(0.51)	(0.49)	(1.87)	(0.067)	0.384
Reading	46.38	46.00	0.376	0.014	
	(0.49)	(0.48)	(1.72)	(0.062)	0.860
SPAG	48.15	47.17	0.98	0.035	
	(0.48)	(0.48)	(1.73)	(0.063)	0.625
Panel B: Cohort 2					
Test Score	45.50	45.55	-0.05	-0.00	
	(0.45)	(0.46)	(1.66)	(0.07)	0.982
Maths	46.87	46.53	0.34	0.012	
	(0.49)	(0.51)	(1.81)	(0.07)	0.892
Reading	44.13	44.58	-0.45	-0.016	
	(0.49)	(0.50)	(1.72)	(0.06)	0.856
SPAG	45.35	45.35	-1.31	-0.047	
	(0.49)	(0.51)	(1.73)	(0.06)	0.566

Notes: This tables shows results of unconditional cross-sectional comparisions, separately for cohorts 1 and 2 (Specification 1 in main text), separately for cohorts 1 (Panel A) and cohort 2 (Panel B). Test Score refers to combined reading and maths tests at age 11. Science scores were only recorded for cohort 1. Number of observations: cohort 1 (cohort 2) 6,436 (6,298). Number of school pairs: Full sample: 90, intended sample: 25, unintended sample 28. Standard errors in parenthesis in column 3 are clustered at school level. Column 5 shows Fisher exact p-values for null effects, based on 10,000 simulations (see Appendix Figures A.2 and A.3.)

Table A.4
Treatment-heterogeneity: Interactions with school characteristics

	(1) Cohort 1	(2)	(3) Cohort 2	(4)
	Main Effect	Interaction	Main Effect	Interaction
Panel A: Ofsted Ratings				
Quality of Teaching	-2.437	0.721	-1.345	0.472
	(5.294)	(2.220)	(5.652)	(2.328)
Behaviour and Safety	1.620	-1.406	-2.767	1.204
	(4.847)	(2.573)	(4.864)	(2.668)
Achievement	-2.634	0.882	-3.644	1.587
	(4.662)	(1.845)	(5.071)	(2.036)
Leadership and Management	-2.814	0.916	-6.908	3.046
	(4.023)	(1.754)	(4.680)	(1.952)
Panel B: School Demographics				
School Size	-5.020	0.117	-4.790	-0.188
	(1.798)	(0.041)	(1.955)	(0.077)
Share Minority	-0.082	1.944	-0.281	4.579
	(1.310)	(3.208)	(1.549)	(3.499)
Share Special Edu. Needs	0.343	-0.484	-0.040	2.940
	(1.303)	(3.524)	(1.467)	(3.949)
Share Free School Meals	1.811	-6.241	0.781	-1.170
	(2.285)	(7.036)	(2.654)	(8.049)

Notes: This tables shows estimates of the effect of the intervention on Test Scores from eight separate regressions in the fulls sample with full controls and pair fixed effects (like Table 3 colum 2), and with additional interactions of the treatment with one school characteristic at a time (left column). Baseline treatment and interaction effects are reported. Ofsted ratings were converted to numbers 1 to 4. Standard errors in parenthesis and are clustered at school level.

Table A.5
Characteristics of Individuals by Treatment Status in Treated Schools

	(1) Treated School	(2) Treated Students	(3) Untreated Students	(4) (2)-(3)	(5) (2)-(3)
Panel A: Cohort 1					
Age-7 Test	15.614	15.584	15.668	-0.083	-0.351
	[3.539]	[3.544]	[3.532]	(0.282)	(0.247)
Share Free School Meals	0.237	0.221	0.268	-0.047	-0.049
	[0.425]	[0.415]	[0.443]	(0.027)	(0.029)
Gender: Male	0.502	0.505	0.496	0.009	0.085
	[0.500]	[0.500]	[0.500]	(0.017)	(0.035)
Share Special Edu. Needs	0.139	0.130	0.157	-0.027	-0.034
	[0.346]	[0.336]	[0.364]	(0.019)	(0.026)
ESL	0.210	0.162	0.300	-0.139	0.010
	[0.408]	[0.368]	[0.459]	(0.083)	(0.015)
Minority	0.240	0.182	0.349	-0.168	-0.003
	[0.427]	[0.386]	[0.477]	(0.087)	(0.016)
Panel B: Cohort 2					
Age-7 Test	15.823	15.870	15.728	0.142	0.362
	[3.455]	[3.385]	[3.593]	(0.288)	(0.429)
Share Free School Meals	0.240	0.235	0.250	-0.015	-0.035
	[0.427]	[0.424]	[0.433]	(0.030)	(0.032)
Gender: Male	0.504	0.513	0.486	0.027	-0.021
	[0.500]	[0.500]	[0.500]	(0.019)	(0.025)
Share Special Edu. Needs	0.126	0.127	0.125	0.002	0.007
	[0.332]	[0.333]	[0.330]	(0.022)	(0.036)
ESL	0.217	0.153	0.346	-0.193	0.028
	[0.412]	[0.360]	[0.476]	(0.081)	(0.028)
Minority	0.240	0.182	0.349	-0.168	-0.003
	[0.429]	[0.368]	[0.492]	(0.085)	(0.020)
Pair FX					X

Notes: Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of labor Economics, 25*(1), 95–135.

Akiba, M., Murata, A., Howard, C. C., & Wilkinson, B. (2019). Lesson study design features for supporting collaborative teacher learning. *Teaching and Teacher Education, 77*, 352–365.

Akiba, M., & Wilkinson, B. (2016). Adopting an international innovation for teacher professional development: State and district approaches to lesson study in florida. *Journal of Teacher Education, 67*(1), 74–93.

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? evidence from matched comparisons in jerusalem public schools. *Journal of labor economics, 19*(2), 343–369.

Barro, R. J. (2001). Human capital and Growth. *American Economic Review, 91*(2), 12–17.

Burgess, S., Rawal, S., & Taylor, E. S. (2021). Teacher peer observation and student test scores: Evidence from a field experiment in english secondary schools. *Journal of Labor Economics, 0*(ja), null. https://doi.org/10.1086/712997.

Carrell, S. E., Hoekstra, M., & Kuka, E. (2018). The long-run effects of disruptive peers. *American Economic Review, 108*(11), 3377–3415.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review, 102*(5), 1805–1831.

DfE. (2013). Reading and maths skills at age 10 and earnings in later life: a brief analysis using the British Cohort Study. *Technical Report*. Department for Education.
DfE. (2016). *Schools, Pupils and Their Characteristics, January 2016*. Dandy Booksellers Limited.
Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of human resources*, 696–727.
for Education, D. (2014). Statistical first release school workforce in england: November 2014.
Fairlie, R. W., & Robinson, J. (2013). Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *American Economic Journal: Applied Economics, 5*(3).211–40
Fernandez, C., Cannon, J., & Chokshi, S. (2003). A us–japan lesson study collaboration reveals critical lenses for examining practice. *Teaching and teacher education, 19*(2), 171–185.
Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., , … Zhu, P., et al. (2011). Middle school mathematics professional development impact study: Findings after the second year of implementation. ncee 2011-4024. *National Center for Education Evaluation and Regional Assistance*.
Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., , … Sepanik, S., et al. (2010). Middle school mathematics professional development impact study: Findings after the first year of implementation. ncee 2010-4009. *National Center for Education Evaluation and Regional Assistance*.
Goodman, S., & Turner, L. (2010). Teacher incentive pay and educational outcomes: Evidence from the nyc bonus program. program on education policy and governance working papers series. pepg 10-07. *Program on Education Policy and Governance, Harvard University*.
Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2).267–71
Hanushek, E. A., & Woessmann, L. (2015). *The Knowledge Capital of Nations*. CESifo Book Series. MIT Press, Cambridge.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics, 95*(7-8), 798–812.
Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics, 1*(4), 85–108.
Jacob, B., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2016). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Technical Report*. National Bureau of Economic Research.
Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics, 26*(1), 101–136.
Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review, 27*(6), 615–631.
Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal, 111*(468), 1–28.
Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review, 99*(5), 1979–2011.
Lewis, C., Perry, R., Hurd, J., & O'Connell, M. (2006). Lesson study comes of age in north america. *Phi Delta Kappan, 88*(4), 273–281.
Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy, 119*(1), 39–77.
Murphy, R., Weinhardt, F., & Wyness, G. (2017). Lesson study evaluation report and executive summary.
Neal, D. (2011). The design of performance pay in education, *4. Handbook of the economics of education* (pp. 495–550). Elsevier.
Papay, J., Tyler, J., & Taylor, E. (2018). Using teacher evaluation data to drive instructional improvement: Evidence from the evaluation paternatship program in tennessee (2015-2020). Unpublished.
Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. (2016). LEARNING JOB SKILLS FROM COLLEAGUES AT WORK: EVIDENCE FROM A FIELD EXPERIMENT USING TEACHER PERFORMANCE DATA. *Technical Report*. NBER Working Paper.
Perry, R. R., & Lewis, C. C. (2009). What is successful adaptation of lesson study in the us? *Journal of Educational Change, 10*(4), 365–391.
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.
Robinson, J. P. (2015). Getting millions to learn: How did japans lesson study program help improve education in zambia?Blog. https://www.brookings.edu/blog/education-plus-development/2015/03/25/getting-millions-to-learn-how-did-japans-lesson-study-program-help-improve-education-in-zambia/.
Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.
Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review, 105*(1), 100–130.
Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D. F., … Stecher, B. M. (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). *Society for Research on Educational Effectiveness*.
Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? experimental evidence from chicago's excellence in teaching project. *Education Finance and Policy, 10*(4), 535–572.
Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7).3628–51
UK, G. (2013). Schools, pupils and their characteristics: January 2014.
Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.