CESifo WORKING PAPERS

8221
2020

April 2020

Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools

Richard Murphy, Felix Weinhardt, Gill Wyness

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8221*

Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools

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, which would have no choice over which teachers would be involved, we find negative impacts of the training (0.1-0.18SD), whereas we find positive impacts in larger schools (0.06-0.17SD). We conclude that the widely-used feedback program that we study is only productive in larger schools, and that centralised one-size-fits-fall teacher training interventions may be harmful.

JEL-Codes: I210, I280, M530.

Keywords: education, teachers, RCT, peer mentoring.

Richard Murphy
University of Texas at Austin
richard.murphy@austin.utexas.edu

Felix Weinhardt
DIW Berlin / Germany
fweinhardt@diw.de

Gill Wyness
UCL Institute of Education / London / UK
g.wyness@ucl.ac.uk

1st April 2020
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 Manheim 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.

1 Introduction

It is well established that teachers are the most important in-school factor in determining student outcomes (Rockoff, 2004; Rivkin et al., 2005). Thus, the fact that there is huge variation in teacher quality (Hanushek and 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 et al., 2014; Hanushek and 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 et al., 2007; Kane et al., 2008), reducing the viability of this solution. An alternative would be to simply dismiss poorly performing teachers (Hanushek and Rivkin, 2010; Chetty et al., 2014), 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 et al., 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 (Lavy, 2009; Goodman and Turner, 2010; Springer et al., 2011; Muralidharan and Sundararaman, 2011; Neal, 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 et al. (2019) 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 and Bruegmann, 2009) as well as from targeted teacher training interventions (Papay et al., 2016).[2]

In this paper, we estimate the causal effect of teacher observation and feedback on student outcomes under RCT conditions by studying 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. (Robinson, 2015; Akiba and Wilkinson, 2016; Perry and Lewis, 2009; Lewis et al., 2006).[3] It consists of a group of teachers planning and observing each others' lessons, and providing feedback as a means to constructively improve their

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

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

[3] The majority of districts in Flordia have mandated the use of Lesson Study (Akiba et al., 2019).

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.[4] 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. We can reject positive effects on student test scores of about ten percent of a standard deviation across all subjects, and effect sizes larger than five percent of a standard deviation in reading and writing tests. However, the overall null finding masks consistently large heterogeneity related to school size. In our study, the schools were responsible for selecting participant teachers into the program, giving rise to possible teacher selection effects in larger schools.[5] Due to the size of the trial, we have sufficient power to examine the importance of this by splitting our sample into small schools (with a single class per grade), and large schools (with multiple classes per grade).

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 0.10 standard deviations (SD) to 0.19SD, in contrast the impact in larger schools increased from improving student outcomes by 0.07SD to 0.17SD. This is consistent with large schools acting optimally and choosing teachers that would gain the most from the program, due to enthusiasm or the need to learn. Unlike Taylor and Tyler (2012) our program did not involve external teacher observers, since our observers were drawn from the internal teacher population. An underlying assumption for peer-to-peer feedback programs to work 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. Our results support this "matching" hypothesis. It is also consistent with the intervention being more disruptive in smaller schools, or teachers being less committed to the program as teachers in smaller schools would not have the option to opt in. An accompanying process evaluation of the program from during the trial found evidence in favor of the matching and disruption mechanisms (indicating that small schools faced more organisational

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

[5]Our school recruitment plan specified that only small (one class per cohort) schools should be recruited to limit teacher selection, but actual recruitment deviated. This is discussed in detail in section 3.5.

challenges and no choice in teacher participants), but not for the enthusiasm channel. We provide additional support that heterogeneity by school size is key, by showing that other school-level factors, i.e. teaching quality, attainment, and school leadership, do not generate heterogeneity in the program effect.

Overall, our findings thus show that this form of teacher feedback is only effective in schools where schools have some choice in which teachers participate. 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. Instead, particular care has to be given to the selection and matching of teachers in situations where good information about teacher quality is not available. On the other hand, 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 and 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. 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 transition 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[6] 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 11 percent of a standard deviation in the year after the evaluation, and about 16 percent of a standard deviation two years later. These effects are in line with our estimates for large primary schools, which are similar in size to American middle schools, but are opposite in sign for small 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, whereas this is more plausible in the Cincinnati setting, where observers were external.

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 0.12 s.d. 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. (2019) 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[7] therefore adding weight to our hypothesis

[6]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.

[7]In 2014 there were 13.3 FTE teachers per primary school, and 64.1 FTE teachers per secondary school (UK,

that 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. 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, with a discussion of our results, and comparison of effect sizes to the ecisting literature, in the concluding section 6.

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. The educational curriculum is organised around Key Stages, where Key Stage 2 incorperates Years 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 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 Year 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. Student are assessed in math and reading and are graded by their teacher, which takes 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 measure for initial student ability.

2013; for Education, 2014)

The age-11 tests examine the students ability in four different areas, maths, reading, Spelling Punctuation and Grammar (SPAG) and science. The first three of these are externally marked on a 100 point scale, which are 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. This is important given the national age-11 assessment changed between the first and second cohorts. The exception to this is Science, which is assessed by the teacher and is only reported in 13 coarse achievment levels which makes it inappropriate to be percentalised. Moreover, there is no science outcome for the second cohort as it was not recorded in 2015/16.

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 and control for school level value added in difference-in-differences specifications. 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. This also reduced testing costs and so allowed us to increase the number of schools the program could be rolled out to on a fixed budget.

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.[8] 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 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

[8]For example Lesson Study Alliance (http://www.lsalliance.org/) helps US teachers, mainly based in Chicago, use Lesson Study; Fernandez et al. (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.

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 were three cycles of the program with each teacher taking the turn to be observed.

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. (2019).

Nevertheless, Lesson Study is a structured program and so the teachers received in-depth training to prepare them for the program. Training consisted of five full training days for teachers participating in 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.

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, Spelling Punctuation and Grammar (SPAG) and science. 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.[9] The program was delivered independently of this impact evaluation by a team at Edge Hill University with support from external consultants.[10]

3.2 Timing of intervention

The trial of the program took place in state primary schools in England[11] 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

[9]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`

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

[11]93 percent of pupils attend state primary schools in England (DfE, 2015)

one year after the end of the intervention, and almost two (cohort 1) or three years (cohort 2) after its start. 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 experiance of the program), are impacted differently.

[Figure 1 goes here]

3.3 Recruitment and teacher selection

The target population for this study are state primary schools in England with above average Free School Meal eligibility (FSM), which stood at 19 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 into the program within 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. 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.[12] 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 0.1s.d.

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.

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.[13]

[12]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.

[13]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.

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.[14]

The recuiting team had difficulty recruiting schools meeting the inital target population requirements of having one class per cohort and above average FSME. In the end, half (91) of the recuited schools had more than one class per cohort and 79 had below 19 percent FSME student population.

The motivation for the focus on one-class-per cohort 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 year groups 4 and 5. In schools with only one class per cohort, this means that both of their teachers from years 4 and 5 had to be selected, with one additional teacher joining from another year group. In contrast, larger schools could meet the requirement of having two teachers of the year groups 4 and 5 included without choosing all of their teachers of these years.[15] As we will argue below, we can exploit the pairwise randimosation that we implemented to estimate causal effects by schools size.

While our data does not allow us to identify the teachers chosen for the intervention, clearly there is a far lower possibility of selection bias arising in the smaller schools, where there is no practical choice over participating teachers in the target grades. To explore the issues relating to teacher selection we split the sample of schools into small schools (our intended sample) and large schools (unitended sample).[16] We define a school to be small if it has 30 of fewer students in each cohort. The reason for this is that 30 is the maximum class size for these Year Groups and our intended sample in the pre-analysis plan was restricted to schools with only one class per cohort. Schools are defined to be large if they have over 30 students in each cohort 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 slightly more

[14]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).

[15]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.

[16]This part of the analysis was not pre-registered. We expected only one-class-per-cohort schools at the time of trial registration.

likely to have Free School Meals (FSM) (22 percent) than pupils nationally (18 percent) or within their LA (19 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 the average attainment at age-7 in these schools is 0.45 levels lower than schools nationally (11 percent of a standard deviation). For the outcomes, age-11 tests, the students perform worse in English by 0.08 levels (8 percent of a standard deviation), but achieve comparably to schools nationally or locally in maths (3 percent of a standard deviation lower). 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.

[Figure 2 goes here] [Table 1 goes here]

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.

Given the power calculations the evaluation had funding to implement the program in 80 schools and therefore the developers we asked to recruit at least 160 schools. Ultimately 182 expressed interest, of which 181 were eligible. There were not the funds to commit to funding the program in half of these schools, therefore 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.[17] When the 83 selected schools were informed that they would be treated, 16 no longer wished to take part, leaving 67 treatment schools.[18] The

[17]The randomization procedure is explained in more detail in Murphy et al. (2017)

[18]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

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.[19] Meaning 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, but we also provide a parallel set of results for the intended (single-form entry) schools and the unintended (multi-form entry) schools. The intended sample consists of 25 pair-IDs where both schools were single form entry. 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 were multi-form entry. We have 27 control schools and 28 treated schools in our unintended sample.

Compliance: Event though we use treatment assignment in our analysis and have outcomes measures from schools that dropped out, it is important to examine if dropout is non-random and potentially affecting any results or generalisations thereof. As explained in Section above the largest deviation from our analysis plan was the fact that half of the recruited schools (91) in the final sample have more than one cohort. Given the potential relevance of school sizes for better teacher matches in the peer-to-peer learning intervention, we therefore present results for the full sample and for pairs of small (intended sample) and for pairs of large schools (unintended sample) separately. This is reflected in the layout of Appendix Table A.1, where we examine dropbout across these three samples. Columns two, four and six report the differences between dropout and stayers, conditional on the pair fixed effects used for the randomisation, for the first (Panel A) and second (Panel B) cohorts. Here, find evidence that dropouts were significantly different than the remaining sample for some characteristics, although no consistent picture emerges comparing the signficant characteristics from the first and second cohorts. Any differences are largely insignificant once we include pair fixed effects. However, for the first cohort, schools that dropped out had slightly higher test scores at age 7, and for the second cohort, schools that dropped out are more likely to be larger, meaning we may potentially be missing out on some positive treatment effects in our results.

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. Individual-level

[19]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

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 are two classes per cohort; the program only involves three teachers and therefore 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. Therefore, the students within a year group that receive treatment might be non-random.

To determine if these excluded classes or new students are systematically different to the treated classes Appendix Table A.2 presents the characteristics of treated and non-treated students within treated schools. Here, we make use of the fact that all treatment schools that did not drop out provided us with lists of students that were taught by teachers in the program. There are almost no differences between treated and untreated students, implying that that classes were not chosen. In the first cohort, treated students are slightly more likely to be male (with pair fixed effects).[20]

As there are some significant observable (and potentially unobservable) differences between the those that were ultimately treated and those who were assigned to treatment (both at the school and student level) and these differences could be correlated with the size of the effect, our main conclusions will be based on intention to treat rather than realized participation. We also present Local Average Treatment Effects (LATEs) results for those schools and students who were actually treated, instrumenting with the assignment status.

[Figure 3 goes here]

3.6 Implementation and fidelity

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

[20] As is expected untreated students come from schools that are signifcantly larger than treated students, because these schools will have a two class entry. However, there is no significant difference in school size when conditioning on pair fixed effects

in five schools, and information on progress provided by four other schools. This provided us with detailed information on the implementation and reception of the Lesson Study program.

While none of the teachers had any experience of the Lesson Study program, some reported having had experience of using classroom observation in the past, 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." Teachers also reported that the requirement to keep personal records of observations was not something they had experienced in the past, and reported that this added to the rigor involved.

The process evaluation concluded that fidelity was high, and schools were found to be implementing the peer-to-peer observation program according to the project design. The intensive 5 day training program may have been responsible for this high fidelity as teachers rated this training highly, referring to it as 'outstanding' or 'high quality'. Many teachers reported that they felt prepared for the program from the outset as the training was well structured, interesting and based on evidence. The importance of teachers observing and not intervening was emphasized particularly strongly during the training, and teachers fully understood the reason for this rule and reported that they followed it. In general there was appreciation for the opportunity to observe colleagues' teaching and learning styles, approaches, and techniques, and to work alongside colleagues with different strengths and expertise. Teachers described how, as the project progressed and a relationship of trust developed, they moved from tentatively advising colleagues to more robust challenges about their teaching and learning practice.

Importantly, the process evaluation found that teachers' behaviour was affected by the program. Two particular aspects were highlighted by participating teachers as particularly beneficial. First, teachers reported the experience of sharing both planning and practice with colleagues was an improvement. Teachers appreciated having dedicated time within the school week to talk to each other about their professional practice, and to share responsibility for planning lessons - an act which was previously done alone. They reported that planning within teams had improved and that team-working more generally was better as a result of Lesson Study.

Second, teachers highlighted the benefits of increased reflection and discussion about pupil learning, and felt the program deepened their level of understanding of how pupils learn. For example, teachers became more aware of those children who go through a whole lesson without participating, or whose vocabulary was too poor to participate. Changes arising from these reflections including teachers giving reticent pupils a waiting time to contribute, rather than moving on when they did not respond, and a school-wide push to improve children's vocabulary.

Regarding the choice of teachers to participate in the program, the process evaluation concluded

that schools used different criteria to select teachers to take part. It 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".[21] In addition to having less choice over which teachers were to be invovled in the prorgam, the process evaluation found that the organisational challenges for planning, delivery and reflection were particularly difficult in smaller schools.

Regardless of school size teachers viewed the program positively after implementing it. They reported finding certain features of the approach useful for their own practice, which reflect the potential mechanisms discussed earlier. First, they found it useful to reflect on their teaching and learning practice and welcomed the opportunity and 'space' within the timetable to reflect on their own practice. Second, they 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 in particular reported positively on the experience of sharing practice with teacher colleagues, shared planning, and identifying complementary skills. The expection from the participants was that the program was a success. We now outline our emprical approach to test this for all schools in our experiment, as well as by school size.

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)[22], 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

$$Y_{ips} = \alpha + \beta D_s + X_{it}{}'\delta + \pi_p + \varepsilon_{ips} \tag{1}$$

where the dependent variable Y_{ips} is the pupil i age-11 test score, in school pair p from school

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

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

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. 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, ethnic minority status and 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.[23]

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, 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 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.[24]

$$T_{is} = \alpha + \beta_1 D_s + X_i{}'\delta_1 + \rho_p + \varepsilon_{ips} \tag{2}$$

$$Y_{ips} = \alpha + \beta_2 \hat{T}_{is} + X_i{}'\delta_2 + \pi_p + \tau_{ips} \tag{3}$$

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

[23]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.

[24]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.

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 treatment effect and the difference between the treatment and control groups 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 school 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 meals receipients (in both our intended and unintended samples). We also observe that schools in our intended sample have an average of 24 students, versus almost 63 students in our unintended sample, which is by construction since the unintended sample consists of pairs of schools with more than one class per grade only. Finally, notice that following our SAP we executed this balancing exercise for a cohort of students *prior* to the student cohorts affected by the intervention, allowing is to show balance in the outcomes measures, the age-11 test scores. This is of paricular interest since in England primary school teachers remain in their year (grade), rather than following cohorts over the years. This balancing result thus shows that there were no significant differences in student intake across treatment and control schools (age-7 test scores), nor in teachers or other factors affecting test score growth between age-7 and age-11 test scores before our intervention.

[Table 2 goes here]

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 pair-wise 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, readings test scores and a score for spelling, and 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 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.

Columns 3 and 4 show results for the intended sample of schools, by estimating effects for schools where both schools of each randomisation pair have only one class per grade. In contrast to the full sample, for this sample of schools we find clear evidence of negative effects of the intervention. For cohort 1, students in schools allocated to the treatment, score on 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 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 or 18 percent of a standard deviation. Moreover, the impacts are statistically significant on 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, math with larger (but not significantly different) impact on maths performance.

These findings are in line with the hypothesis that teacher observation is more effective in larger schools. The negative effect observed in smaller schools, on the other hand, may be evidence of the distraction effect associated with running such a programme, or that the lack of opt-in on behalf of teachers negatively impacts outcomes.[25]

[Table 3 goes here]

[25]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.

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. We now present Local Average Treatment Effects (LATE) by instrumenting actual participation status with school assignment. Columns 1 to 3 of Table 4 instrument participation at the school level, using the randomized assignment as instrument. Columns 4 to 6 repeat this exercise but instead instrument for student-level participation. The latter is defined by student-level lists the schools sent to us. 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 new information. First, the school LATE is larger than that of the student LATE, which reflects the fact that not all students in a school participate in the program. Second, there is no difference between the first stages using school assignment between intended and unintended samples, establishing there is no differential attrition between small and large schools. Third, for the student LATE 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 year group 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 sample and the positive estimates for the unintended samples are confirmed, 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 for 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 it depends on the setting.

[Table 4 goes here]

5.4 Heterogeneity

5.4.1 School-level heterogeneity

To recap, we document causal effects of the opposite sign across small and large schools. This is just one school characteristic. To investigate if the effectiveness of the progam varies by other school level characteristics, we now present estimates for the impact of the program on four different school-level characteristics. The four school-level characteristics that we can do this for come from government (Ofsted) inspections of the schools, where schools are rated by inspectors in terms of 1) Quality of School Leadership, 2) Teaching Quality, 3) Safety and Behaviour, and 4) Pupil Attainment. Using this data, we have categorised the Ofsted-ratings into high (Outstanding, Good) or low (Satisfactory, Inadequate). We then estimate the ITT effects of the intervention on our four main student-level outcomes, separately for pairs in which both schools have the high rating and again for pairs in which both schools have the low rating.

For neither the first or second cohort does splitting the sample by any of these four characteristics generate significant effects on average test scores, unlike school size. However there is a consistent pattern that the coefficients are postive (negative) when teaching and leadership quality is low (high). We infer from these findings that key school characteristics such as leadership, teaching quality, discipline, and achievement are not strongly correlated with program effectiveness.

In conclusion, we believe this additional evidence on school-level heterogeneity of the effects of the intervention shows no clear candidates that could drive the heterogeneity in the effects that we have documented across pairs of different school sizes above.

[Table 5 goes here]

5.4.2 Student-level heterogeneity

Table 6 presents student level heterogeneity ITT analysis, interacting treatment status with five binary student characteristics. While student-level heterogeneity cannot explain the differences in the program effects across small and large schools as students are similarly represented in both types of schools, it is important in its own right in better understanding potential workings of teacher feedback programs on student learning outcomes. The five characteristics analysed are students who are eligible for free school meals (FSME), speak English as additional language (ESL), belong to an ethnic minority, are low achievers in terms of their age-7 outcomes, or are male.

[Table 6 goes here]

Out of the forty interaction terms estimated here, two are statistically significant at the five percent level (or higher). In cohort 1, the overall effect on the SPAG outcome is not significant, but the interaction for minority students is positive and statistically significant from zero. In cohort 2, there is no overall effect on math, but boys are negatively affected relative to girls. Given the number of coefficents tested we would expect this number of coefficents to be signficant at the 5 percent level, and so we conclude that there is little eveidence for student-level heterogeneity.

6 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 key finding is that, whilst overall we find no significant impact of the programme, we uncover heterogeneity of the impact across schools. In smaller schools where no teacher selection was possible since there was one class per cohort, the program had a clear negative impact on student test scores. By contrast, in larger schools where there was more opportunity for teacher selection, there were gains of 7 percent of a standard deviation for the first cohort and 17 percent for the second.[26] There are three potential reasons for this. First is that there is a greater probabilty of there being variation in teacher quality in a large school from which the other teachers can learn. Second, we have qualitative evidence that coordination of teachers observing and provide feedback is more difficult in smaller schools. Third, teacher enthusiasm is potentially dampened in smaller schools by not having the option to opt into the program. Evidence from the process evaluation found evidence in favour of the first two of these mechanisms, but not for the third. 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, may be ineffective, and may even lead to negative impacts. Moreover we find that a key determinant of effectivness is school size, with no other headline school-level characteristics (such as achievment, leadership or disipline) showing evidence for heterogeneity of the treatment.

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 11.2 percent of a standard deviation in maths achievement in the first year after the observations, and effects of 15.8 percent of a standard deviation two years after the observations. [27]. However, for

[26]This might also explain differences in findings to Papay et al. (2016) who find positive co-worker effects among pairs of teachers who were purposefully paired up based on previous effectiveness measures.

[27]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

our full sample we can where we can reject effects of up to 11.04 for cohort 1 and 12.62 percent of a standard deviation cohort 2.[28]. 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. Our positive finding in large schools is also in line with recent evidence from Burgess et al. (2019) 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 are likely generalisable since they are based on a large sample of primary schools, and provide an evaluation of Lesson Study, which is widespread and being used in over fifty countries. 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.

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

References

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

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

Akiba, M. and B. Wilkinson (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. and V. Lavy (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., S. Rawall, and E. S. Taylor (2019). Teacher peer observation and student test scores: Evidence from a field experiment in english secondary schools. Technical report, Working Paper). Cambridge, MA. Retrieved from https://scholar. harvard. edu ....

Chetty, R., J. N. Friedman, and J. E. Rockoff (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.

Fernandez, C., J. Cannon, and S. Chokshi (2003). A us–japan lesson study collaboration reveals critical lenses for examining practice. *Teaching and teacher education 19*(2), 171–185.

for Education, D. (2014). Statistical first release school workforce in england: November 2014.

Garet, M. S., A. J. Wayne, F. Stancavage, J. Taylor, M. Eaton, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, 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., A. J. Wayne, F. Stancavage, J. Taylor, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, S. Sepanik, 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. and L. Turner (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. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review 100*(2), 267–71.

Hanushek, E. A. and L. Woessmann (2015). *The Knowledge Capital of Nations*. CESifo Book Series. MIT Press, Cambridge.

Harris, D. N. and T. R. Sass (2011). Teacher training, teacher quality and student achievement. *Journal of public economics 95*(7-8), 798–812.

Jackson, C. K. and E. Bruegmann (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., J. E. Rockoff, E. S. Taylor, B. Lindy, and R. Rosen (2016). Teacher applicant hiring and teacher performance: Evidence from dc public schools. Technical report, National Bureau of Economic Research.

Jacob, B. A. and L. Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics 26*(1), 101–136.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review 27*(6), 615–631.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review 99*(5), 1979–2011.

Lewis, C., R. Perry, J. Hurd, and M. O'Connell (2006). Lesson study comes of age in north america. *Phi Delta Kappan 88*(4), 273–281.

Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from india. *Journal of political Economy 119*(1), 39–77.

Murphy, R., F. Weinhardt, and G. Wyness (2017). Lesson study evaluation report and executive summary.

Neal, D. (2011). The design of performance pay in education. In *Handbook of the Economics of Education*, Volume 4, pp. 495–550. Elsevier.

Papay, J., J. Tyler, and E. Taylor (2018). Using teacher evaluation data to drive instructional improvement: Evidence from the evaluation paternatship program in tennessee (2015-2020). Unpublished.

Papay, J. P., E. S. Taylor, J. H. Tyler, and M. Laski (2016). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. Technical Report 21986, NBER Working Paper.

Perry, R. R. and C. C. Lewis (2009). What is successful adaptation of lesson study in the us? *Journal of Educational Change 10*(4), 365–391.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Robinson, J. P. (2015, March). Getting millions to learn: How did japans lesson study program help improve education in zambia? Blog.

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., D. Ballou, L. Hamilton, V.-N. Le, J. Lockwood, D. F. McCaffrey, M. Pepper, and B. M. Stecher (2011). Teacher pay for performance: Experimental evidence from the project on incentives in teaching (point). *Society for Research on Educational Effectiveness*.

Steinberg, M. P. and L. Sartain (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. and J. H. Tyler (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., S. Sexton, J. Mulhern, D. Keeling, J. Schunck, A. Palcisco, and K. Morgan (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

Tables and figures

Figure 1: Timeline of intervention

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**

Notes: Red square shows treatment period and cohorts.

Figure 2: Treatment and control schools

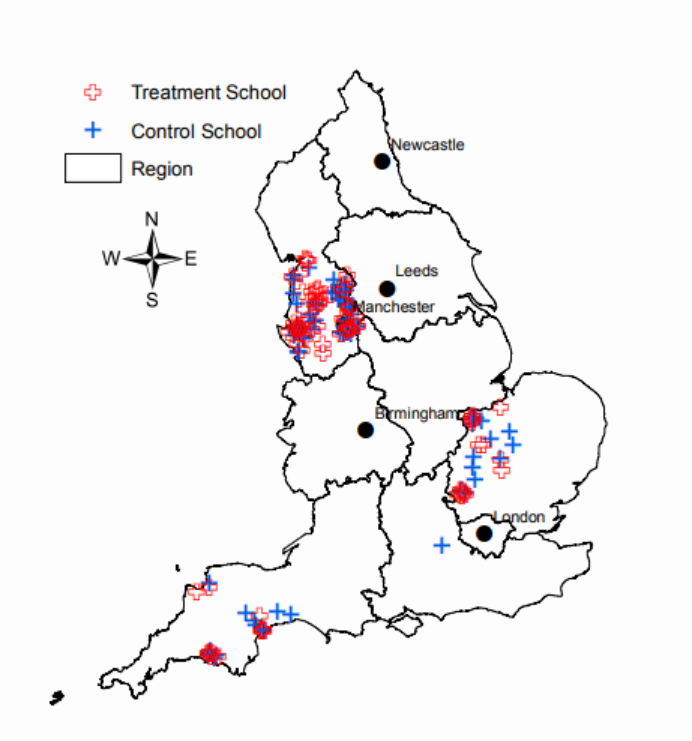


Figure 3: Consort flow diagram

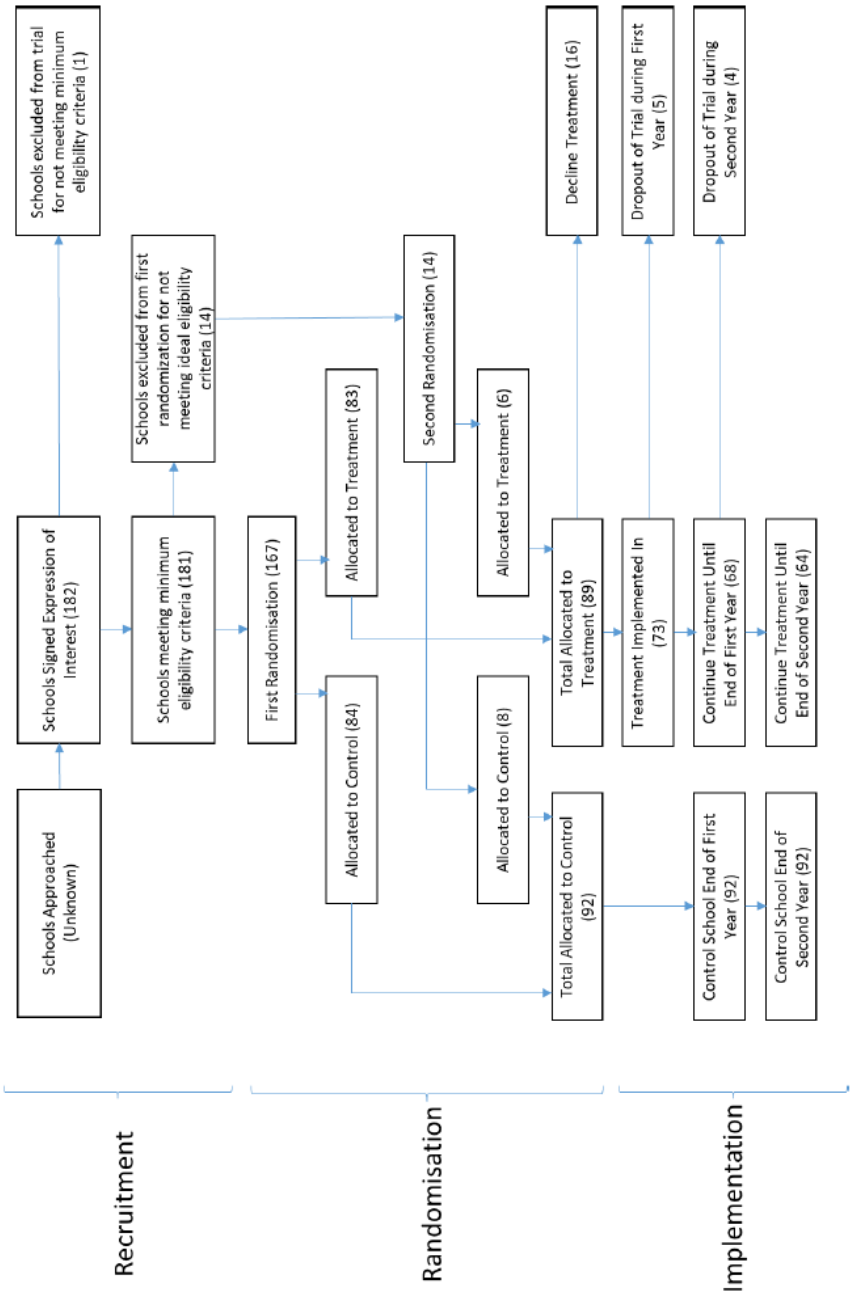


Table 1: National and local representativeness of sample

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]
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.

Table 2: Randomisation tests: cohort 1 and cohort 2

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample (all)		Intended		Unintended	
	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). Number of school pairs: Full sample: 90, intended sample: 25, unintended sample 28. Standard deviations of variables shown in square parenthesis, standard errors clustered at the school level shown in round parenthesis.

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.

Table 4: IV analysis

	(1)	(2)	(3)	(4)	(5)	(6)
	School LATE			Student LATE		
	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.

Table 5: Heterogeneity - School Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores		Maths		Reading		SPAG	
	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)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores		Maths		Reading		SPAG	
	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction	Main Effect	Interaction
Panel A: Cohort 1								
FSME	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.

Appendix

Appendix 1: Participating Local Authorities

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

Tables and Figures

Table A.1: Analysis of School-Level Dropout

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample (all)		Intended		Unintended	
	Dropout	Difference	Dropout	Difference	Dropout	Difference
Panel A: Cohort 1						
Age-7 Test	15.455	0.560	15.785	0.491	15.930	0.318
	[3.372]	(0.238)	[3.654]	(0.557)	[3.563]	(0.174)
Share Free School Meals	0.261	-0.002	0.169	-0.117	0.234	0.032
	[0.439]	(0.030)	[0.376]	(0.040)	[0.423]	(0.021)
Gender: Male	0.514	0.019	0.485	-0.021	0.499	-0.004
	[0.500]	(0.026)	[0.502]	(0.055)	[0.500]	(0.016)
Share Special Edu. Needs	0.141	0.004	0.162	-0.012	0.119	0.010
	[0.349]	(0.018)	[0.369]	(0.041)	[0.324]	(0.010)
School Size	51.619	10.951	23.238	3.924	62.240	0.731
	[22.032]	(4.371)	[4.756]	(2.064)	[25.295]	(4.304)
Panel B: Cohort 2						
Age-7 Test	15.612	0.039	16.482	1.420	15.833	-0.187
	[3.384]	(0.292)	[3.469]	(0.376)	[3.575]	(0.211)
Share Free School Meals	0.247	0.023	0.202	-0.069	0.254	0.070
	[0.432]	(0.028)	[0.403]	(0.059)	[0.435]	(0.022)
Gender: Male	0.480	-0.057	0.470	-0.089	0.493	-0.019
	[0.500]	(0.024)	[0.501]	(0.032)	[0.500]	(0.015)
Share Special Edu. Needs	0.132	0.024	0.101	0.036	0.126	-0.017
	[0.338]	(0.023)	[0.302]	(0.033)	[0.332]	(0.019)
School Size	50.759	9.332	24.940	1.234	62.501	-4.992
	[21.904]	(4.782)	[4.526]	(2.251)	[23.433]	(8.026)
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: Analysis of Individual-Level Dropout 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)
School Size	46.896	43.329	53.508	-10.179	0.002
	[24.641]	[23.041]	[26.112]	(4.793)	(0.002)
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)
School Size	47.422	43.863	54.618	-10.755	0.000
	[23.257]	[20.308]	[26.903]	(5.586)	(0.000)
Pair FX					X

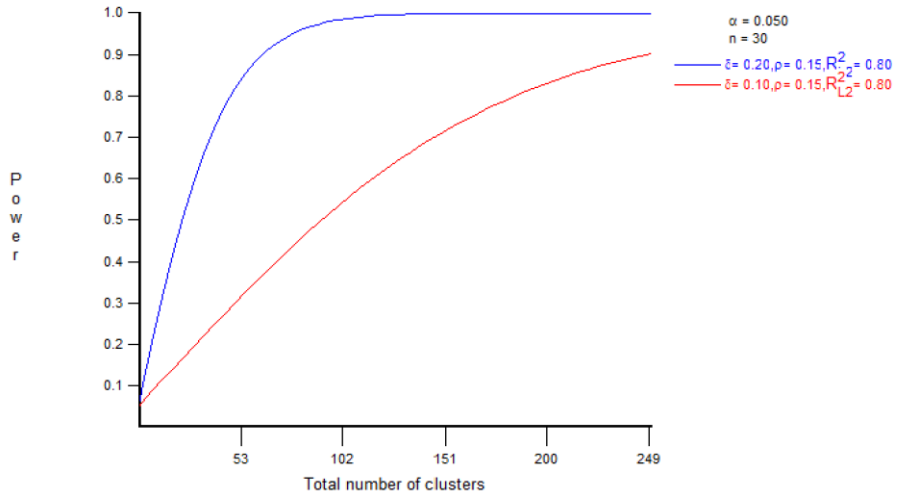
Notes: Standard deviations of variables shown in square parenthesis in columns 1-3. Standard errors clustered at the school level shown in round parenthesis in columns 4-5.

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

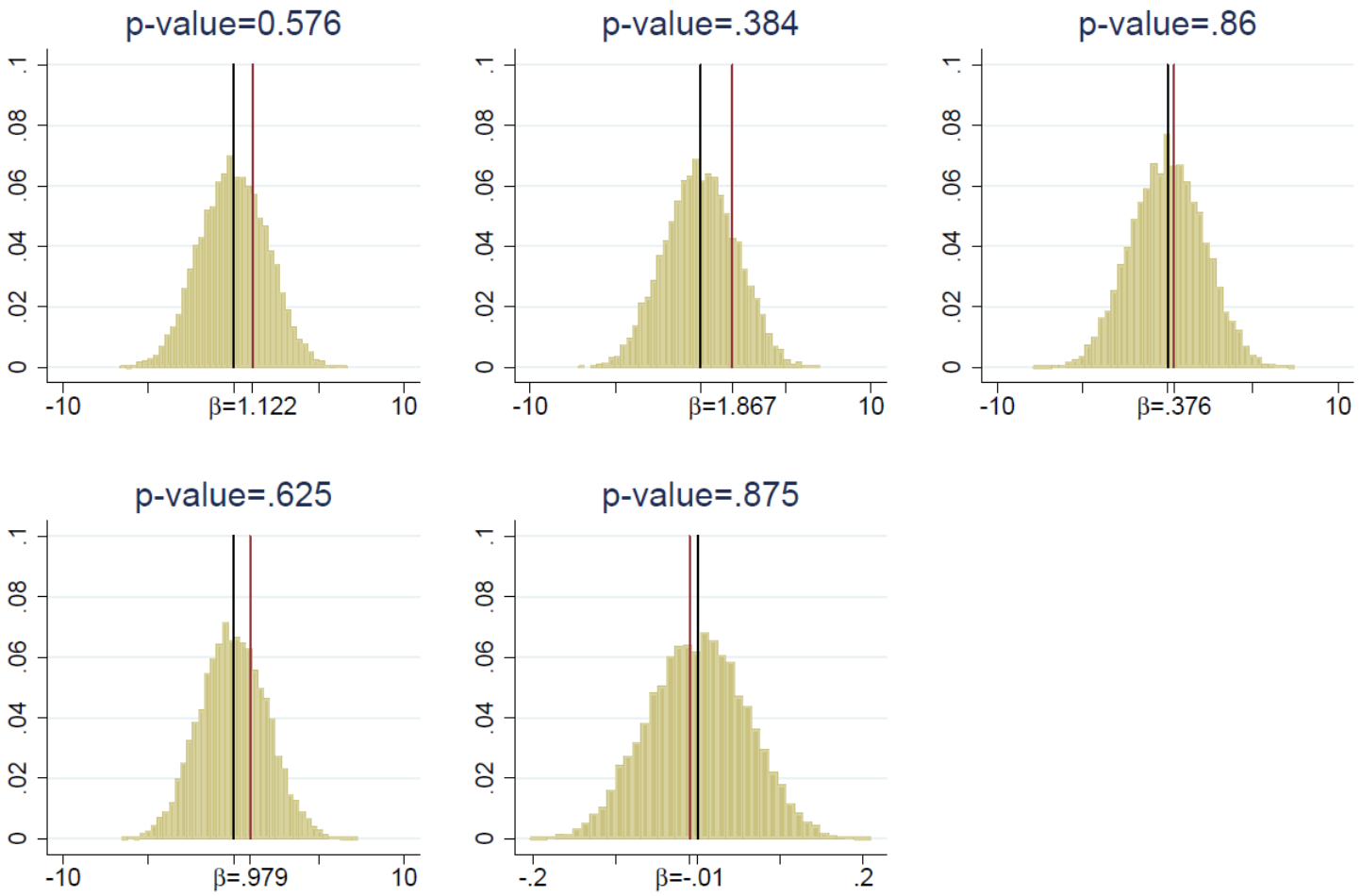
Notes: This tables shows results of unconditional cross-sectional comparisions, separately for cohorts 1 and 2 (Specification 1 in main text), separately for cohorts 1 (Panel A) and cohort 2 (Panel B). Test Score refers to combined reading and maths tests at age 11. Science scores were only recorded 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.)

Figure A.1: Power calculations, pre-trial



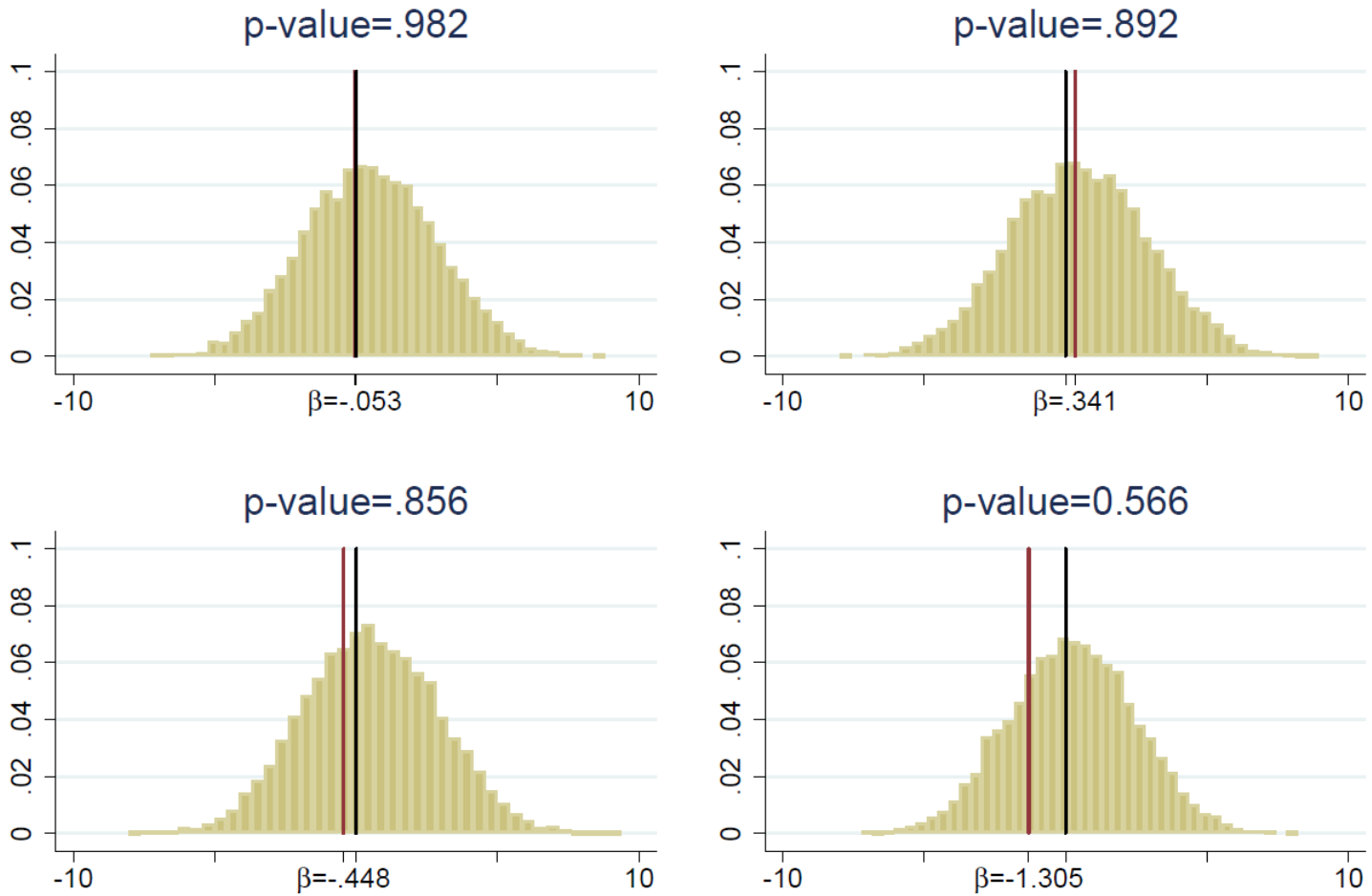
Notes: Blue line indicates power with effect size of 0.2 s.d., red line effect size of 0.1 s.d.

Figure 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.

Figure 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.