PROJECT TITLE	School Partnership Programme
DEVELOPER	The Education Development Trust
EVALUATOR	UCL Institute of Education
PRINCIPAL INVESTIGATOR(S)	Jake Anders (Impact Evaluation), David Godfrey (Implementation and Process Evaluation)
PROTOCOL AUTHOR(S)	Jake Anders, John Jerrim, Louise Stoll, Toby Greany, David Godfrey
STUDY DESIGN	School-level matched difference in differences
AGE RANGE	10-11
NUMBER OF SCHOOLS	437
NUMBER OF PUPILS	N/A
PRIMARY OUTCOME	KS2 SATS maths performance
PROTOCOL VERSION	4.0

Protocol version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	24 May 2018	N/A
2.0	25 February 2019	Addition of appendix reporting matching process
3.0	1 August 2020	Revisions due to disruption caused by COVID-19. Addition of Theory of Change model as appendix. Ultimately unpublished due to further COVID-19 disruption.
4.0	1 May 2021	Revisions due to renewed disruption caused by COVID-19. These identify impact evaluation outputs that are no longer possible.

Table of contents

Intervention

The Education Development Trust (EDT)'s School Partnership Programme (SPP) is a structured approach to cluster-based school collaboration, through the provision of a coherent and consistent approach to peer review that aims to drive improvement across all schools involved in the cluster. The programme aims to develop a culture of partnership working through school self-evaluation, peer review and school-to-school support.

SPP is a peer review model that is intended to build capacity and capability across clusters so they can gradually take more responsibility for their own development and maturity, and lead their own improvement. Over time, local areas will own the SPP model, and continue to develop it so it has impact locally. SPP provides frameworks and tools, training and professional support, and is designed to incorporate and build on, not side line, schools' existing best practice.

Significance

The need for lateral school-to school partnerships has become apparent in the face of evidence that neither top-down centrally imposed change, nor pure competition can achieve sustained improvement across school systems (Burns and Koster, 2016). The aim, rather, has been to 'unleash greatness' by asking school system leaders to work together in ways which transfer knowledge, expertise and capacity within and between schools, so that all schools improve and all children achieve their potential (DfE, 2010).

This has implications for accountability, with the drive for a 'self-improving school system' leading to an increase in engagement in peer evaluation to promote self-accountability (Greany and Higham, 2018). This is seen as a key step towards schools self-regulation in which schools take greater ownership of their quality assurance, not only through self-evaluation but through exposing their work to the scrutiny and perceptions of trusted peers (Matthews and Ehren, 2017). This accords with the outcomes of an international comprehensive survey of assessment and evaluation in 28 countries by the OECD (2013). In finding little evidence of peer review, the OECD report's authors identified developing school evaluation capacity as a priority, proposing that school leadership teams collaborate to identify common challenges and devise common approaches to peer evaluation.

In March 2020, delivery of SPP was disrupted by COVID-19 restrictions. As a result, the decision was taken to extend supported delivery of the programme in schools that are part of this evaluation from two to three academic years, to allow participating clusters of schools successfully to complete the aims of the intervention in a way that COVID-19 restrictions prevented them from doing. This was to provide a chance for some schools to complete the number of planned review cycles while for others, provide a chance to add a further cycle. Further changes to the model, involve moving all workshops and training online and the development of a rapid review model, so that schools can conduct peer review visits partially or completely online. This will mean that year 3 will be a different experience while retaining the key principles and logic model.

Further details of the model and its hypothesised logic model, which informs the evaluation design, are available in Appendix C.

Methods

Research questions

The primary objective of this evaluation was to estimate the effect of participating in the EDT School Partnerships Programme (SPP) for two years on pupils' attainment. **Due to cancellation of KS2 National Curriculum tests in the summer of 2020 and the summer of 2021 it will not be possible for us to provide evidence on this objective.** When KS2 National Curriculum tests were initially cancelled for summer 2020 the possibility of using

summer 2021 tests as our post-intervention outcome of interest was explored, but the subsequent cancellation of these tests removed this option.

In addition, the impact evaluation planned to answer the following secondary research questions, **which will also not be possible for the same reason**:

1. Does participating in EDT SPP have an effect on the attainment of young people ever identified as eligible for free school meals (FSM)?

Due to the two lockdowns, first in spring 2020 and then again in 2021, many of the IPE activities were curtailed. These re-started again from summer term 2021 and take into account EEF support for continued delivery of SPP until end of Autumn term 2021. The implementation and process evaluation questions have been modified to reflect the lack of impact pupil attainment data (see above) and subsequently seek to answer the following questions (these are now ALL the research questions and thus are re-numbered from 1):

1. In what ways does the SPP influence the capability, culture and practice of partnerships, leadership and teachers in involved schools?
2. In what ways do the elements in the School Partnership Programme theory of change work in achieving participants' perceived forms of impact?
3. What factors influence schools' and clusters' ability to engage in, participate fully in and successfully implement and sustain their involvement in the programme?
4. What distinguishes schools and clusters that have not continued with the programme?
5. What difference has COVID-19 made to the operation, participant engagement and perceived forms of impact of SPP (i.e. RQs above)?

Design

This evaluation was designed as an embedded mixed methods evaluation, incorporating a school-level matched comparison difference in differences impact evaluation, with an Implementation and Process Evaluation (IPE). Due to cancellation of KS2 SATS in the summer of 2020 and the summer of 2021 it now consists only of the (adapted) IPE.

The approach to the impact evaluation was chosen because a randomised controlled trial would not have been feasible with this programme, partly due to the scale required (because of the grouping of schools into clusters) and because of the difficulty of forming schools into clusters while expecting them not to cooperate in the case that they are allocated to a control group.

Participants

Target recruitment was 50 clusters of English state-funded primary schools to be recruited with an approximate cluster size of 6, making 300 schools in total, with the proviso that if cluster size is smaller than expected additional recruitment would be undertaken to bring the number of schools recruited up to 300. In the event, the project team (EDT) successfully recruited far more schools than anticipated, providing a sample of 437 English state-funded primary schools in 85 clusters (average cluster size of just over 5). All recruited schools receive the intervention as part of this project, while statistical matching methods is used to identify the counterfactual group.

In order to be considered for participation, schools had to agree to cooperate with the project and evaluation teams during the trial (further details of these requirements are outlined in the School Partnerships Programme Agreement, between EDT and each partnership, a template for which is included with this document in Appendix A).

The project team advertised the trial and also approached schools through their existing networks. Where possible it aimed to recruit schools that have larger populations of individuals receiving FSM.

Impact evaluation

It will not be possible to estimate and report impact estimates due to the cancellation of the planned outcome measures in 2-3 years post-treatment, as discussed above. Instead, we will provide a report of the position of the evaluation at the pseudo-randomisation date, incorporating reporting of matching and pre-treatment trends in outcome measures, prior attainment measures and related variables (to the extent possible given changes in attainment measures during this period). An early version of this reporting is available as Appendix 2 to the protocol since version 2. From this we will draw lessons for future research considering adopting a similar approach.

Outcome measures
The primary outcome of interest was school average performance in KS2 maths tests (mat_average). More specifically, given that this is a difference in differences design, the outcome will be the difference in differences between treatment and matched comparison schools in pre- and post-treatment years.

The secondary outcome of interest is school average performance in KS2 reading tests (read_average). The same specifics apply as for the primary outcome above.

Additional secondary outcomes will be the outcomes above for the FSM sub-groups within schools. These will be recovered using school-level variables available for this purpose (mat_average_fsm6cla1a and read_average_fsm6cla1a).

None of these outcome measures are now available in the pre-specified post-treatment year (or the subsequent year, which was initially explored as an alternative).

Matching
For clarity, matching was carried out on the basis of a 1:1 nearest neighbour propensity score matching without replacement, including exact matching on key characteristics (likely to include school type and government office region), application of a caliper[1] on the propensity score matching, and imposition of common support. Since clusters are observed in the treatment group but not in the pool of potential matched comparators, it is necessary to match at school-, rather than cluster-, level; the importance of the clustering of treatment schools was planned to be recognised in the analysis. The matching exercise was carried out following the pseudo-randomisation date, with the final treated and matched samples considered fixed at this point, before any outcomes data were expected to be available.

Characteristics that were assessed for inclusion in matching (including interactions and higher order polynomials of these terms) were:

- Number of pupils
- Attainment in school measured by KS2 average points score in each previous year 2010-15
- Prior attainment of intake measured by KS1 average points score in each previous year 2010-15
- School type (academy vs. other)
- Ofsted rating
- IDACI quintile
- Geographical location (longitude, latitude)
- Local authority area
- Government office region

[1] We will explore the quality of fit of differing widths for our preferred strategy but plan to start this search with a caliper of 0.2 of the standard deviation of the logit of the propensity score (as recommended by Austin, 2011). Note that we will explore the robustness to varying this caliper, as discussed below.

The identification of pseudo-clusters among untreated schools based primarily on geographic characteristics, to which treated schools could be matched, was explored but rejected. We made this decision as it resulted in worse balance of school-level characteristics between the treatment and matched comparison groups in a pilot exercise. Instead, geographic characteristics (e.g. latitude, longitude, government office region, IDACI quintile) were included as part of the matching process itself.

Difference in differences

While matching attempts to ensure that the treated and comparison group are comparable on the basis of observable characteristics, there is still the potential for confounding due to unobservable differences between the treated and matched comparison schools. We planned to take a difference in differences approach to deal with remaining time-invariant unobservable characteristics. This means that our impact estimate makes the assumption of common trends i.e. that in the absence of the treatment the change in our outcomes of interest between the pre- and post-treatment period would have been the same between our treatment and our matched comparison schools (Anders et al. 2017, ch. 4). We aimed to improve the plausibility of this assumption by using previous years' average attainment as part of our matching.

We planned to compare the differences in outcomes between treated and matched comparison schools in 2017/18 (pre-treatment) with the difference in outcomes between treated and matched comparison schools in 2019/20 (post-treatment) – two years after the intervention began. **This will not now be possible due to cancellation of KS2 SATS in 2019/20 (and 2020/21).**

Sample size calculations

We conducted our sample size calculation for the KS2 maths outcome, since this was the primary outcome of interest. Sample size calculations were based on an estimated Minimum Detectable Effect Size (MDES) of 0.20 and the following assumptions: power of 0.8 for a two-tailed 0.05 significance test, treatment assignment at cluster-level, an intra-cluster correlation of 0.10[2] and 6 schools within each cluster.

In conducting this calculation, we assumed that 0.40 of post-test variance at school- and 0.70 at cluster-level is explained by the pre-test and lagged performance (in the setting of a difference in differences this is based on variation explained by lagged performance in the outcome variable and, in this case, performance a KS1 "pre-test"). The pre-test/post-test correlation assumptions are based on estimates derived from a database of schools previously treated by EDT.[3]

These requirements suggested a requirement of approximately 300 treated schools with the final average cluster size not exceeding 6 (as this would reduce the power). Based on discussions with the project team and EEF at project set-up, this was set as the recruitment target.

Since all analyses planned to use school-level variables, the power calculation for average performance of FSM pupils is no different to that for the overall outcome.[4]

[2] It is difficult to choose an ICC value in this setting given that little evidence exists for intra-cluster correlations at school-cluster (rather than within school) level. As a result, we choose 0.10 as being at the lower level of within school ICCs found by EEF in previous trials, based on an assumption that within-cluster variance is likely to be higher than within-school variance.

[3] Specifically, we ran a school-level model of average points score in 2014 on average points score for the same cohort at KS1 and average points score in 2013 in the same school, allowing for cluster-level variance components. This estimated within-cluster variance explained at 0.42 and between-cluster variance explained at 0.70.

[4] We note the risk of figures among the FSM sample being suppressed in schools where there are 3 or fewer pupils who are eligible for FSM.

Analysis plan
The following analysis will not be carried out for lack of outcome measures.

However, we planned to carry out the impact evaluation analysis as follows, estimating the effect of the intervention using a linear model on school-level data from the pre- and post-treatment periods (as defined above). Raw outcome variables from the NPD, as described in the outcome measures section above, were to be used in all models. Cluster-level clustered standard errors were to be calculated in order to take into account the potential dependence of the results among school clusters; schools in the matched comparison group were to be treated as independent from one another for the purposes of calculating standard errors.

The model was to include a treatment indicator, a post-treatment period indicator, an interaction term between the treatment indicator and the post-treatment period indicator, and school average performance at Key Stage 1 (tks1average, or an updated version of this) as an additional way to reduce bias in the estimator (Imbens & Rubin, 2015, ch.18) i.e. as follows:

$$mat\_average_{it} = \alpha + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 Treat_i * Post_t + tks1average_{it} + \varepsilon_{it}$$

As this was be estimated on the treatment sample defined at the pseudo-randomisation date, the coefficient on the interaction term (β_3) would have recovered the Intention to Treat (ITT) Average Treatment on the Treated (ATT) estimate of impact.

We were to calculate Hedge's g effect size by dividing this coefficient by an estimate of the unconditional pooled total variance of the outcome variable and applying the appropriate correction factor. 95% confidence intervals were to be estimated by inputting the upper and lower confidence limits of the coefficient from the regression model into the effect size formula.

An estimate of the intra-cluster correlation of the outcome measure was to be extracted by estimating a variance components model for this purpose.

As noted above, the regression model would have included a pre-test variable in order to improve the precision of the estimates.

We were to estimate the impact on average performance of FSMever pupils using a separate model using the relevant outcome variable from the NPD. This was to be carried out for both maths and English performance.

We were to estimate treatment effects for compliers (both "minimal" and "optimal") at both school-level and cluster-level using a sub-group analysis defined by a school-level and cluster-level measures of compliance with the intervention (the cluster-level measure is based on an aggregation of the school-level measure). The definition of these fidelity measures is discussed as part of the Implementation and Process Evaluation Stage 4 below.

Robustness checks
We planned a battery of robustness checks of both the matching and the difference in difference elements of the design in order to to establish the credibility of the estimates. These were planned to include:

Matching
- Selection of two nearest neighbours;
- Varying the caliper width (including half and double of the caliper selected for our preferred approach);
- Exclusion of items from the matching equation;
- Removal of exact matching characteristics;

- Removal of imposition of common support;
- Use of kernel matching as an alternative to nearest neighbour matching.

We also planned to explore the balance of baseline characteristics in the matched models these produce in order to select 5 well-matched alternative specifications to use as the core robustness check models for impact estimation.

Difference in differences

- Use of two years prior to implantation as baseline (rather than one year prior to implementation);
- Specification of the estimation model as a fixed effects estimation rather than difference in differences.

Implementation and process evaluation methods

Original overview:

The purpose of the process evaluation is to establish fidelity and to assess the factors which affect impact from the different phases of the SPP on the stakeholders within the project, and which may explain the findings of the quantitative evaluation. We will also look for evidence of wider issues which may need to be considered in any further roll out of this programme and other whole school-level interventions. The process evaluation will involve the following:

- UCL IOE attendance at EDT training events
- Baseline and final interview surveys of 437 head teachers in intervention schools
- UCL IOE interviews of school leaders and teachers in all schools in 2 clusters
- UCL IOE attendance at localised cluster pre-review training, reviews, and follow up (3 schools)
- UCL IOE interviews of school leaders and teachers at 8 matched schools
- Interview surveys of 437 matched school head teachers

Revised rationale due to Covid-19:

RQ1 remains unchanged and tests the SPP theory of change (see appendix C). To reflect the lack of reference to literacy/numeracy outcomes, instead we will be reviewing *perceptions of impact* as reported in surveys and interviews at pupil, leadership, teacher, school and partnership levels (RQ2). We know from the autumn workshops that reviews for many partnerships have switched to areas such as well-being, the recovery curriculum and online learning too, so these can be explored. We have also distinguished research questions (3 and 4) that compare those schools that stayed in all the way through the programme to those that left earlier. We have added group interviews to capture cross-cutting data on the role of partnership-leads, improvement champions and associates (the ones that lead the partnership training and workshops). Given the extended period of delivery to include autumn term 2021, the end surveys will now be conducted after the last 'review of reviews' workshops (EDT to confirm exact date, late Autumn term 2021). Numbers of participating schools will be lower than those stated above, however we hope to capture data from schools that dropped out to compensate for this. There will thus be three end surveys of head teachers: i) schools that remained in the programme until autumn 2021, those that left the programme earlier iii) matched schools. This will delay the draft report until Spring 2022 but will also allow us to use all other data to better inform the construction of the final surveys. The review of reviews sessions will also be an ideal platform for EDT and the IPE team to encourage participation in the QA telephone survey. Below are revised and additional IPE activities under original stages and dates:

Stage 1 (approx. May/July 2018):

Baseline telephone Interview surveys
To gather data from all treatment schools, we propose that telephone baseline (followed by final surveys in the last stage) of intervention schools be carried out. This would include information on "business as usual" and differences between "business as usual" and the intervention and gain a wider view of fidelity and/or impact as measured qualitatively.

Stakeholder Interviews
We will carry out individual telephone interviews with a number of key stakeholders as agreed with the delivery partners to explore intentions around the nature and reach of activity, programme differentiation.

Revised:

Additional stakeholder interviews of partnership leads, SPP associates and improvement champions in summer term 2021

Stage 2 (approx. starting date March 2018 and various dates depending on training arrangements – also ongoing throughout two years)

Observing EDT peer review training events
Members of our team with expertise and knowledge of leadership development and peer review will lead the observations and fieldwork. The IPE team will attend and observe all types of training sessions delivered by the training provider, as well as reviewing the materials used.

A sample of EDT and partnership training will be observed, including:

1. Training of reviewers – Senior Leaders and Improvement champion training (Year 1 - Spring 2018; Autumn 2018; Year 2 – Autumn 2019).
2. Baseline, interim and summative impact workshops led by EDT (Year 1 - Spring 2018, Autumn 2018, Summer 2019; Year 2 – Autumn 2019, Spring 2020; Summer 2020).
3. Further training for middle leaders and governors to start Year 2 cycle (Autumn 2019)

Ongoing desk review of related SPP materials (Summer 2018 – Summer 2020)
Working closely with the delivery partners (EDT), we will aim to draw on data collected by them and school-level data where possible. For example, we will seek permission from all participating schools to be given access to their self-evaluation and peer-review framework and tools, online audit tools and school level documentation.

Revised:

We will sample all additional training and workshops, including Partnership Lead forum (June 2021) and review of reviews workshops for case study schools (see below) November 2021.

Stage 3: Detailed case studies of the intervention and matched comparison groups (Autumn 2018 and Autumn 2019).

Case studies from intervention group
We will follow two case study clusters and their intervention schools (5 schools per cluster) over the two years of the intervention. Initial data on school and cluster characteristics gathered as part of the quantitative evaluation will be used to inform selection of this sample. The clusters will be located in different contexts and at different stages of development to explore potential developmental changes. We will conduct one visit per year per school to interview people (a minimum of 3 per school e.g. head teacher, involved senior or middle leader, Improvement Champion and another teacher), using instruments based on the logic

model. We will also carry out interviews at impact workshops to monitor the peer-review process: their own review at their school, their attendance at a local peer review, follow-up school-to-school support or other relevant cluster activity (*dates as above*). Schools in these case study schools will also share with us documentation used in their peer review activities. To further inform the case studies, a sample of reviews will be observed, including:

1. Pre-conversations (Year 1 – Summer 2018; Year 2 – Summer 2019)
2. Review visit (Year 1 – Summer 2018; Year 2 – Summer 2019)
3. Follow-up improvement workshop (Year 1 – Summer 2018; Year 2 – Summer 2019)

Case studies from matched comparison clusters
We will follow two case study clusters of matched schools (4 schools per cluster) over the two years, also using information on school and cluster characteristics gathered for the quantitative evaluation. We will also conduct face-to-face interviews with head teachers and senior leaders in two matched clusters of schools to probe any similar interventions in which they might be involved, focusing on similarities and differences in the process, engagement of teachers and wider impact on pupils. This will help us understand how 'standard' school practices in the areas of self-evaluation, peer review and school to school support compare with the SPP model.

Revised

We will shadow 3 further reviews, either from case study groups and/or elsewhere, to capture new adaptations to the peer review process and new areas of focus.
We will complete interviews of case study school staff that we were unable to do due to Covid-19 interruptions.
We will conduct further interviews with partnership leads of both case study clusters
We will conduct interviews in four matched schools, using longer interviews and with the secondary purpose of piloting some questions for end telephone surveys

Stage 4 (Summer 2020):

Final telephone interview surveys of intervention schools and telephone interviews of matched comparison schools
The final telephone surveys in the last stage of intervention schools would include information on fidelity, dosage, quality, reach, responsiveness, programme differentiation, adaptation. Some Likert scale responses which would be included in the baseline survey would compare distance travelled in terms of levels of trust, openness, support and challenge.

Revised

End surveys of: i) schools that remained in the programme until autumn 2021, those that left the programme earlier iii) matched schools in late autumn 2021.

In addition, we would provide descriptions of non-compliant, partially-compliant and compliant schools, based on objective data recorded centrally by EDT (see table below). This would allow for sub-group analysis in the impact evaluation looking at how effects vary by compliance. Using only EDT data has the advantage of not having to factor-in non-completion of surveys into our judgements about compliance. We would also overcome issues with self-reporting of issues to do with responsiveness, reach, programme differentiation, etc.

However, separately to the impact evaluation, we will also conduct further analysis as part of our IPE, based on additional data from our survey on other aspects of (self-reported) fidelity to the SPP.

School-level compliance categories and criteria for School Partnership Programme

We have clarified the school-level compliance criteria in order to remove ambiguity, particularly in the context of disrupted and extended delivery due to COVID-19 restrictions. Beyond simple clarification, we have varied the review visits criterion to be considered a fully compliant school in order not to exclude schools whose year 2 visits were disrupted by COVID-19 restrictions, but successfully resume these in year 3, and to ensure that those who did conduct visits in year 2 continue these in year 3 to be considered fully compliant.

Categories	**Attendance to training/workshops**	**Review visits**
Non-compliant schools	No attendance to delivered training sessions/workshops or have otherwise indicated that they have dropped out of the programme.	No review visits *or* have otherwise indicated that they have dropped out of the programme
Minimally compliant schools	Attendance to less than 75% of delivered training sessions/workshops	Has hosted *at least* 1 peer review visit across programme
Fully compliant schools	Attendance to 75% or more of delivered training sessions/workshops	Has hosted *at least* 2 review visits across programme (if they *had not* completed 2 before COVID-19 lockdown) or *at least* 3 (if they *had* completed 2 before COVID-19 lockdown)
Source of data	EDT database	EDT database

In order to be categorised as minimally or fully compliant both “Attendance to training/workshops” and “Review visits” criteria must be met.

Cluster-level compliance categories and criteria for School Partnership Programme

Cluster-level compliance categories are based on aggregation of school-level compliance of all schools within a cluster (as identified at the beginning of the study, regardless of cluster reconfiguration). This aggregation is carried out as follows:

Minimally compliant: A minimally compliant cluster contains *no* non-compliant schools and at least one fully compliant school.

Fully compliant: A fully compliant cluster contains a *maximum of one* school that is partially compliant and *the rest* are fully compliant.

All clusters that fail to meet *either of these criteria* will be considered to be **Non-compliant**.

Ethics and registration

Ethical approval has been sought following UCL Institute of Education staff ethics approval procedure. It was approved on 20 March 2018.

This protocol is has been registered at www.controlled-trials.com, and the assigned International Standard Randomised Controlled Trial Number (ISRCTN) is ISRCTN20687346.

Personnel

Project team
Jenni Rolls, John Cronin, Anne Cameron, Maggie Farrar (SPP)

Evaluation team
Jake Anders, John Jerrim, Louise Stoll, David Godfrey (UCL), Toby Greany (Nottingham)

The teams will have the following roles within the evaluation:

Design of the trial
- Sample size calculation – Evaluation team
- Refinement of matching approach – Evaluation team

Delivery of the intervention
- Recruitment of schools – Project team
- Delivery of intervention – Project team

Measurement of outcomes
- Collection of outcomes data from administrative sources – Evaluation team

Impact analysis – Evaluation team
Qualitative analysis – Evaluation team

Risks

The data security policy of UCL is available at https://www.ucl.ac.uk/informationsecurity/policy/public-policy/information-security-policy.pdf.

Some of the key risks are summarised in the table below:

Issue/risk	**Risk level**	**Action to address issue/reduce risk**
Dropout / non-compliance of settings	Medium	We want to avoid attrition of schools from the project as much as possible. We plan to minimise attrition by ensuring that schools that sign up are committed (by asking them to sign a Memorandum of Understanding). Keeping them informed of progress and providing reminders of next steps will be important for retention. The project team should also monitor changes in key personnel to ensure ongoing commitment. Minimising the data collection burden on schools will also be important for retention. We will also randomise only after schools have followed consent collection procedures, provided the necessary student data.
Difficulty recruiting schools	Low to medium	We are confident that the project team will convey the importance of the evaluation to settings and the value to them of taking part. To understand whether recruited settings are atypical in some way (which would affect external validity), we ask that the project team keep records of settings approached and, where possible, of reasons for not participating.
Missing outcome data	Low	We are confident that we should be able to recover estimated outcomes for all treated schools. **Note:** Ultimately, this risk did come to pass given the cancellation of KS2 SATS in our expected outcome year and the subsequent year. However, we continue to believe that our estimation of this risk was reasonable at the time given the unprecedented nature of the disruption. Furthermore, the changed circumstances have affected out estimation of this risk going forward. We think there is now a **moderate** risk of missing outcome data in the coming years due to a combination of further public health (e.g. COVID-19 school closures), central political (e.g. cancellation of SATS), and local political (e.g. SATS boycotts) risks. (This reassessment was

		carried out in light of initial COVID-19 disruption and proved prescient given the second year of cancellation.)
Poor balance between treatment and matched comparison group	Medium	We have a large pool of donor schools which should mean it is possible to get a well-balanced matched comparison group. Furthermore, we do not rely solely on balance on observables to justify the credibility of our treatment estimates.
Treatment variation	Medium	We view this not so much as a risk but as the reality of implementing such an intervention. The impact estimates (Intention to Treat) therefore relate more to the type of treatment likely to prevail in practice rather than the type of impact that could be seen were it possible to achieve laboratory-type conditions. Nevertheless, understanding treatment variation is important and will be a key focus of the process study.
Unexpected absence or loss of team members	Low	The team will substitute for each other during any short-term absence. In the event of longer periods of unplanned absence or departure, we will recruit replacements. UCL have other experts in evaluation and education who could substitute for members of the team, should this be necessary.

Timeline

Date	Activity
October 2017 – February 2018	**Recruitment (Project team)**
October 2017 – February 2018	**Pre-Randomisation Data Collection (Project team)**
March 2018	**Pseudo-randomisation date (Evaluation team)**
September 2018 – December 2021	**Intervention carried out (Project team)**
October 2020	**Release of originally planned outcomes data (KS2 tests in Summer 2020) by DfE (Evaluation team). *This has been cancelled.***
March 2018 – December 2021	**Implementation and Process Evaluation Fieldwork (Evaluation team)**
October 2021	**Release of initially considered outcomes data (KS2 tests in Summer 2021) by DfE (Evaluation team). *This has been cancelled.***
December 2021– April 2022	**Report Writing (Evaluation team)**

References

Anders, J., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., Groot, A., Sanders, M., & Allen, R. (2017) Evaluation of Complex Whole-School Interventions: Methodological and Practical Considerations. Report to the Education Endowment Foundation.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat, 10(2), 150-161. doi:10.1002/pst.433

Burns, T. and Köster F. (eds.) (2016), *Governing Education in a Complex World*, Educational Research and Innovation, OECD Publishing, Paris.

DfE (2010), *The Importance of Teaching: The Schools White Paper*, Cm 7980, Department for Education, London.

Greany, T., and Higham, R., (2018) *Hierarchy, Markets, Networks and Leadership: analysing the 'self-improving school-led system' agenda in England*. IOE Press: London.

Matthews, P, and Ehren, M, (2017) 'Accountability and Improvement in Self-improving School Systems'. In Greany T, and Earley P, (Eds) School Leadership and Education System Reform, Bloomsbury: London.

OECD (2013), Synergies for Better Learning: An International Perspective on Evaluation and Assessment. PISA, Paris: OECD.

Appendix A: Schools Partnership Programme Agreement

Agreement between:

Education Development Trust of [Highbridge House, 16-18 Duke Street, Reading, RG1 4RU] "Education Development Trust"

and {insert name of partnership}"Partnership"

This agreement is for the period: 1st January 2018 to 30th August 2020 ("The Initial Period")

- The Partnership has the right to terminate this agreement at the end of the Initial Period by giving 3 months' notice. Such notice must be signed by an appropriate representative of the Partnership.

Operational expectations of each party:

Partnership:

1. The Partnership will ensure that it is represented at all relevant meetings and events unless there is an unavoidable reason to prevent this.
2. The Partnership will nominate a Partnership lead. This individual will hold the post for duration of the contract (unless the post is delegated to another representation of the cluster) and will be responsible for:
   a) providing leadership to the Partnership, ensuring that Partnership members work effectively together and that all members participate in all aspects of the SPP, including Self-Review, Peer Review, School-to-School Support and developing an Action Plan
   b) being the key point of contact between the partnership and Education Development Trust, providing information and data to Education Development Trust as requested;
   c) agreeing to participate in evaluation activities, with other members of the partnership e.g. Improvement Champions, with the independent evaluation team at UCL Institute of Education, including interviews.
3. All members of the Partnership will agree to share and be transparent on data, in accordance with the Data Sharing Agreement with other members of their Partnership, Education Development Trust, and the independent evaluation team at UCL Institute of Education;
4. The Partnership will maintain confidentiality: any information, data or documents received by any member of the Partnership will not be shared with any third party, or used outside of the SPP and the independent evaluation team at UCL Institute of Education, without the consent of the disclosing party;
5. In order to support system improvement, the Partnership will share the outcomes their peer reviews with the rest of their partnership, Education Development Trust and the independent evaluation team at UCL Institute of Education. Parties will not share these outcomes without the permission of the school and reviewers concerned and will use them only for the purposes of the SPP and the independent evaluation by UCL Institute of Education;
6. In order to support governors' accountability, all members of the Partnership agree to ensure that the governing body of each school is kept fully informed of their Partnership work and receives relevant feedback following peer reviews. Individual partnerships will be expected to discuss and agree the governance arrangements for their Partnership;
7. The Partnership will develop their own Memorandum of Understanding to encapsulate their partnership approach but also to ensure that clear lines of accountability are agreed.
8. If any concerns about a school come to light following a peer review, such as evidence of illegal activity or safeguarding concerns, the Partnership will ensure that the appropriate body is informed;
9. The Partnership will undertake to appoint the appropriate personnel, including Improvement Champions prior to initial training.
10. Partnership members will pay Education Development Trust the agreed financial contribution for involvement in the project.

Education Development Trust:

1. Education Development Trust will assign the Partnership a named Education Development Trust contact who be their key point of contact for the project.
2. Education Development Trust will retain ownership of all intellectual property in materials created or used by Education Development Trust and/or a Partnership for the purposes of the SPP. Education Development Trust will allow all Partnerships to use all materials for the sole purpose of the SPP only while they are affiliated to the programme;
3. Education Development Trust will maintain confidentiality: any information, data or documents received by Education Development Trust will not be shared with any third party, or used outside of the SPP and the independent evaluation team at UCL Institute of Education, without the consent of the disclosing party;
4. Education Development Trust will ensure that it keeps partnerships informed of its research programme and international opportunities, insofar as the same are relevant and appropriate to the Partnership or the SPP.

General legal terms

1. Nothing in this agreement is intended to create or shall constitute or be deemed to constitute a legal partnership or joint venture between the parties.
2. This agreement is not intended to confer any rights upon any third parties or persons not a party to it.
3. This agreement shall form the entire understanding between Education Development Trust and the Partnership and may only be amended by written agreement of either party.
4. This agreement shall be governed by the laws of England and Wales.

Data Sharing

Effective partnerships will work collaboratively in a mutually challenging and supportive way in order to bring about system-wide improvement. To be truly effective, partnerships will benefit from sharing their data in an open and transparent way, allowing data to be used effectively to underpin the work of peer review teams. In order to achieve this, members are asked to sign a data sharing agreement.

Permission to share data

I agree to share school and subject level data with the schools indicated below on the terms outlined by the Partnership below:

The Partnership recognises that the practice of data sharing must comply with all confidentiality, data protection, intellectual property and safeguarding regulations of the schools and contained within legislation.

The Partnership understands that any data shared should not contain any pupil information, must be treated confidentially and used to support collaborative work and the sharing of best practice by schools within the group.

Where data is shared with Education Development Trust, the terms of this agreement will be upheld by Education Development Trust.

SPP EEF trial Phase 1 programme training and support (March 2018 - July 2019)

Peer review & improvement champion training x 1 event - for heads, deputies and nominated improvement champions (improvement champions are a shared capacity across the cluster – up to 2 places per school for Senior Leaders (i.e. the Headteacher, Deputy etc) who will be undertaking the peer reviews. A further 3 places across the cluster for the role of the improvement champion will also be allocated.

Improvement Champion (IC) training day 2 – for nominated improvement Champions (those identified above)

Impact / review workshops x 3 over the year - to establish and monitor impact through baseline / interim & summative improvement conversations with SPP Associate.

Access to all tools; SPP handbook / improvement framework etc.

SPP EEF trial Phase 2 programme and support (September 2019 – July 2020)

Peer Review Training for middle leaders / beyond x 1 event – 2 places per school

Collaborative leadership training for senior leaders x 1 event – 2 places per school

Impact / review workshops x 3 throughout the year - to establish and monitor impact through baseline / interim & summative improvement conversations with SPP Associate.

Throughout the duration of the programme, clusters will have full affiliation to Education Development Trust (including access to events, termly newsletters)

Costs of the programme

The total cost of the programme is £1300 +VAT per primary school participating.

The cost for other phases of schools is £2600+VAT.

The Partnership lead will coordinate the payment from each school, although each school can make their payment directly to Education Development Trust.

Full payment must be made 30 days after being invoiced by Education Development Trust

Partnership details

Please confirm details the name of your 'Partnership Lead' and the names of the 3 'Improvement Champions' you have nominated for your cluster.

Name or Partnership Lead	• **Name 1 / email address**
Names of Peer Reviewers (x 2 per school)	• **School 1 – names / email address** • **School 2 – names / email address** • **School 3 – names / email address**

	• School 4 – names / email address • School 5 – names / email address • School 6 – names / email address • School 7 – names / email address
Names of Improvement Champions (x 3 per cluster)	• Name 1 / email address • Name 2 / email address • Name 3 / email address

Signed on behalf of each school in the Partnership below:

	Name of school and address	LA establishment number	Name of headteacher	Headteacher email address	Amount paid to Education Development Trust by school	Name and email for contact to be invoiced	Signed an (please i scanned ve your sign unfortuna are una accept 't signatu
1							
2							
3							
4							
5							
6							
7							

Add more rows if required.

Signed on behalf of Education Development Trust:

Dated:

EDT SPP Evaluation Matching Exercise

As per the evaluation protocol, we have conducted a matching exercise to identify a preferred matched comparison sample to use in our difference in differences estimation of the impact of the Schools Partnership Programme (SPP).

Unmatched sample

We begin by documenting the balance between treatment groups and all other primary schools in the regions of England where recruitment occurred (all except West Midlands and North East). Schools in Opportunity Areas were also excluded given the other work going on in these local authority districts. This included exclusion of a small number of treated schools who had been recruited in Opportunity Areas.

Table 1. Unmatched sample

Characteristic	Treated	Untreated	Std. Diff
KS2 Reading Score in 2017	104.9	104.3	0.18
KS2 Maths Score in 2017	104.4	104.2	0.07
KS2 Reading Score (FSM)	101.7	101.6	0.03
KS2 Maths Score (FSM)	101.4	101.7	-0.12
KS1 Intake Score in 2017	16.1	15.9	0.15
Academy	0.19	0.22	-0.08
IDACI Quintile 1	0.20	0.18	0.05
IDACI Quintile 2	0.25	0.16	0.21
IDACI Quintile 3	0.25	0.29	-0.09
IDACI Quintile 4	0.19	0.18	0.03
Ofsted: Outstanding	0.17	0.15	0.04
Ofsted: Good	0.71	0.62	0.18
Ofsted: RI	0.10	0.20	-0.28
Mean Abs. Std. Diff.			0.12
Treated N			383
Untreated N			11428

Notes. Reporting mean characteristics in treated and untreated schools within areas in which recruitment occurred. “Std. Diff” = Standard differences calculated by dividing means by overall sample standard deviation. “Mean Abs. Std. Diff” = Mean absolute standard difference calculated across characteristics in table. IDACI Quintile 5 and Ofsted: Inadequate categories excluded since these are determined by the remainder of the other categories of this variable. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

There is significant imbalance on many characteristics, although when we begin modelling it is notable that quite a few of these characteristics end up not being statistically significant predictors. It is also important to highlight that level differences in balance do not in themselves invalidate the assumptions inherent in a difference in differences method. Instead, the appropriate assumption is of common trends, the plausibility of which we will discuss later.

Preferred specification

In our preferred specification, matches are found using a nearest neighbour algorithm with no replacement (in practice, allowing replacement makes no difference in this application, seemingly because there are plenty of potential matched comparators available) using the MatchIt package in R, based on a treatment propensity score estimated from the following generalised linear model with a logistic link function:

$$logit(Treat_i) = \beta_0 + \beta_1 Read_{i2017} + \beta_2 Maths_{i2017} + \beta_3 KS1_{i2017} + \beta_4 PupilNo_{i2017} + \beta_5 PupilNo^2_{i2017} + \beta_6 PupilNo^3_{i2017} + \beta_7 Acad_{i2017} + \beta' \boldsymbol{Ofsted_{i2017}} + \beta' \boldsymbol{IDACI_{i2017}} + \beta' \boldsymbol{Region_i} + \varepsilon_i$$

where $Treat_i$ is our 0/1 indicator of whether school i is participating in the SPP project, $Read_{i2017}$ is average KS2 reading score of the school in 2017, $Maths_{i2017}$ is average KS2 maths score of the school in 2017, $KS1_{i2017}$ is average KS1 score of the school’s intake (among those taking KS2 tests in 2017), $PupilNo_{i2017}$ is the number of pupils in the school in 2017 (with higher orders of this variable in the following two terms), $Acad_{i2017}$ is whether the school is an academy in 2017, $\boldsymbol{Ofsted_{i2017}}$ is a vector of binary variables indicating school’s most recent Ofsted rating in 2017, $\boldsymbol{IDACI_{i2017}}$ is a vector of binary variables indicating the quintile group into which the school falls in terms of the average Index of Deprivating Affecting Children and Infants (IDACI) of its intake, $\boldsymbol{Region_{i2017}}$ is a vector of binary variables indicating the government office region in which a school is located, and ε_i is an idiosyncratic error term.

This model was based on iterative testing of model fit of the matching variables proposed in the evaluation protocol with an important exception. In the evaluation protocol, we proposed the potential use of lagged performance variables in order to improve the probability of achieving common trends in our matched sample. However, further reading has suggested that, while this would appear to improve the plausibility of this assumption, in fact it may cause problems with regression towards the mean in the future trends that we will use to estimate the treatment effect.

1:1 nearest neighbour matching based on the estimated propensity score was carried out without replacement and also enforced exact matching on the school's IDACI quintile, on urban/rural classification, and on school region. Exact matching on Ofsted rating was explored but rejected as reduced the sample of schools that could be matched without offering an obvious benefit (see details of alternative specifications). We used a caliper of 0.2 in line with the advice of Austin (2011). Schools outside the range of common support were excluded, although when we tested removing this restriction it only resulted in one treatment school being excluded.

Alternative specifications

We tried a number of alternative specifications as part of the matching process, a selection of which are reported here.

- Caliper of 0.1
    - Narrowing the caliper which should, other factors equal, result in increased exclusion of worse matches but reduce sample representativeness. Based on our analysis of balance (reported below) the 0.2 caliper was retained as this alternatives did not obviously perform better.
- Caliper of 0.4
    - Widening the caliper which should, other factors equal, result in reduced exclusion of worse matches but improve sample representativeness. Based on our analysis of balance (reported below) the 0.2 caliper was retained as the alternatives did not obviously perform better.
- Remove imposition of common support
    - Common support imposition only resulted in the exclusion of just one treated school, so this made very little difference to our matched sample.
- Matching with replacement
    - Allowing matching with replacement made no difference to our matched sample. The nearest neighbour algorithm always chose a different comparison school for each treated school even when this constraint was relaxed.
- 2 Nearest Neighbours
    - This made a small improvement to the average balance of the matched sample. However, we do not adopt it as our preferred specification as this improvement is small and the the 1:1 matching was specifically chosen in the protocol for clarity.
- Leave out KS1 intake measure
    - KS1 intake was not a statistically significant predictor of being in the treatment group, nevertheless we retain it in our model in the preferred specification. Removing this seems to worsen balance overall.
- No exact matching
    - All exact matching characteristics are removed (but retained as categorical predictors in the propensity score model). This worsened overall balance.
- Exact matching also on Ofsted judgement
    - This worsened average balance (including on attainment measures) other than for the Ofsted judgement itself.
- Generalised Additive Modelling of school location replacing exact matching on urban/rural classification
    - Urban/rural classification was removed as an exact matching criterion and empirically determined interacted polynomials capturing variation in schools' longitude and latitude were instead used to predict treatment status. This resulted in a closer match in geographical distribution but worsened average balance overall and reduced sample size.

Table 2. Balance characteristics by matching methods

Characteristic	0.1 Cal	0.2 Cal	0.4 Cal	2NN	No Ex.	No KS1	Ex. Of.	No CS	GAM	Unmatch
KS2 Reading Score	0.02	0.02	-0.02	0.00	0.05	0.01	0.02	-0.04	-0.04	0.18
KS2 Maths Score	0.06	-0.01	-0.04	-0.01	0.01	-0.02	0.01	-0.11	-0.04	0.07
KS2 Reading Score (FSM)	-0.06	-0.02	-0.06	-0.05	0.00	-0.05	-0.08	-0.14	-0.08	0.03
KS2 Maths Score (FSM)	-0.05	-0.06	-0.05	-0.07	-0.15	-0.13	-0.07	-0.21	-0.12	-0.12
KS1 Intake Score	0.05	0.02	-0.02	0.03	0.08	0.07	-0.03	0.08	-0.05	0.15
Academy	-0.03	-0.04	-0.05	-0.08	-0.04	-0.04	-0.06	-0.06	0.10	-0.08
IDACI Quintile 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
IDACI Quintile 2	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.21
IDACI Quintile 3	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00	0.00	-0.09
IDACI Quintile 4	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.03
Ofsted: Outstanding	0.01	-0.04	-0.01	0.00	-0.01	-0.08	0.00	-0.02	-0.01	0.04
Ofsted: Good	0.01	0.03	0.02	0.00	0.01	0.04	0.00	0.03	0.01	0.18
Ofsted: RI	-0.02	-0.01	-0.05	-0.02	0.01	0.02	0.00	-0.02	0.00	-0.28
Mean Abs. Std. Diff.	0.02	0.02	0.02	0.02	0.04	0.04	0.02	0.05	0.03	0.12
Treated N	367	374	374	374	374	374	366	375	362	383
Untreated N	367	374	374	743	374	374	366	375	362	11428

Notes. Reporting "Std. Diff" between treated and comparison schools identified by each matching method described. Standard differences calculated by dividing means by overall sample standard deviation. "Mean Abs. Std. Diff" = Mean absolute standard difference calculated across characteristics in table. IDACI Quintile 5 and Ofsted: Inadequate categories excluded since these are determined by the remainder of the other categories of this variable. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

Based on analysis of how these alternative specifications in terms of balance and differences from the preferred specification, and in line with the evaluation protocol, we chose five on which will estimate impact estimates as robustness checks:

- Caliper of 0.4
- 2 Nearest Neighbours
- No exact matching
- Exact matching also on Ofsted judgement
- Generalised Additive Modelling of school location replacing exact matching on urban/rural classification

In the remainder of this paper, we provide details of the matched sample balance and common trends in the preferred specification.

Balance

When we restrict our comparison group sample using the matching process documented above, the balance on these characteristics improves substantially and is now no greater than 0.05 standard deviations for all the characteristics we consider. We view this as particularly important for the measures of attainment, although we note that our causal identification strategy does not specifically rely on no differences at baseline, since our treatment estimate is based on differences in differences, rather than simple differences.

Table 3. Balance statistics in the preferred sample compared to in the unmatched sample

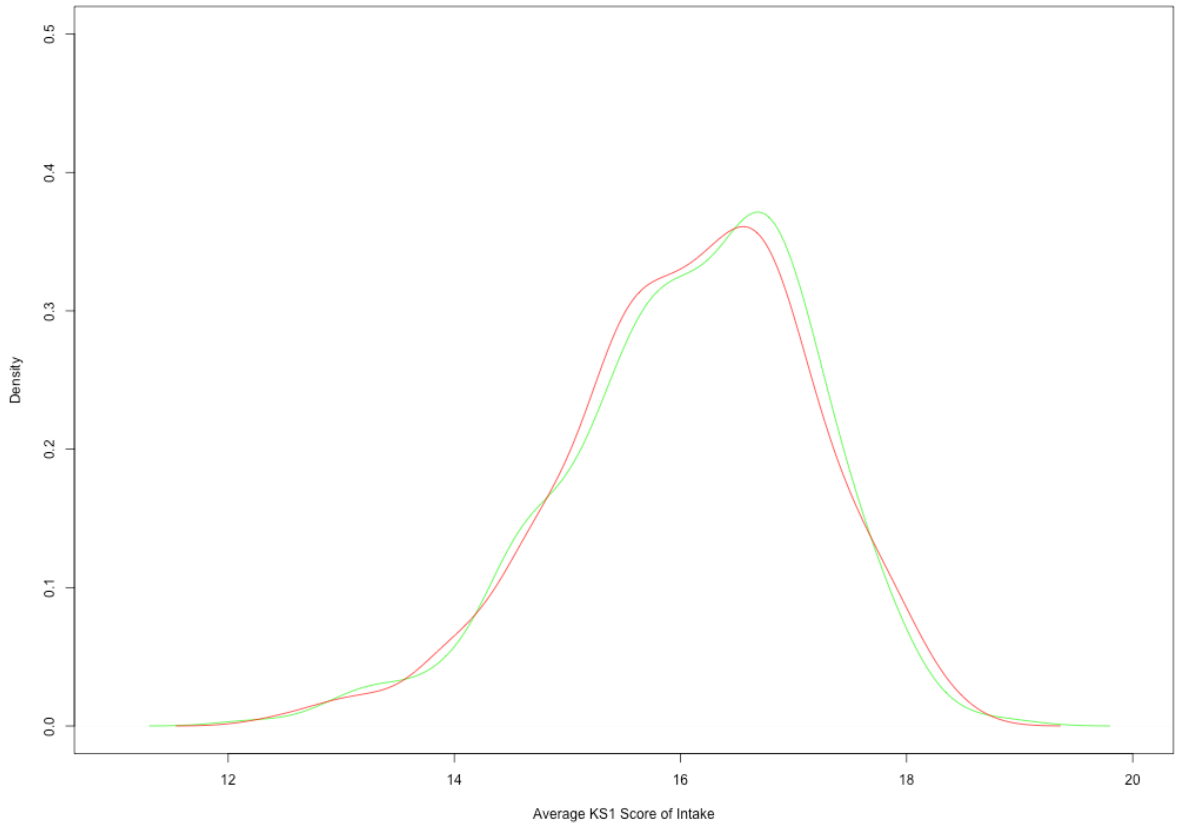
	Matched			Unmatched		
Characteristic	**Treated**	**Untreated**	**Std. Diff**	**Treated**	**Untreated**	**Std. Diff**
KS2 Reading Score in 2017	104.9	104.8	0.02	104.9	104.3	0.18
KS2 Maths Score in 2017	104.4	104.5	-0.01	104.4	104.2	0.07
KS2 Reading Score (FSM)	101.7	101.8	-0.02	101.7	101.6	0.03
KS2 Maths Score (FSM)	101.4	101.6	-0.06	101.4	101.7	-0.12
KS1 Intake Score in 2017	16.1	16.1	0.02	16.1	15.9	0.15
Academy	0.18	0.20	-0.04	0.19	0.22	-0.08
IDACI Quintile 1	0.21	0.21	0.00	0.20	0.18	0.05
IDACI Quintile 2	0.24	0.24	0.00	0.25	0.16	0.21
IDACI Quintile 3	0.24	0.24	0.00	0.25	0.29	-0.09
IDACI Quintile 4	0.19	0.19	0.00	0.19	0.18	0.03
Ofsted: Outstanding	0.17	0.19	-0.04	0.17	0.15	0.04
Ofsted: Good	0.71	0.70	0.03	0.71	0.62	0.18
Ofsted: RI	0.10	0.10	-0.01	0.10	0.20	-0.28
Mean Abs. Std. Diff.			0.02		0.12	
Treated N			374		383	
Untreated N			374		11428	

Notes. Reporting mean characteristics in treated and untreated schools in preferred matched sample ("Matched") and within areas in which recruitment occurred ("Unmatched"). "Std. Diff" = Standard differences calculated by dividing means by overall sample standard deviation. "Mean Abs. Std. Diff" = Mean absolute standard difference calculated across characteristics in table. IDACI Quintile 5 and Ofsted: Inadequate categories excluded since these are determined by the remainder of the other

categories of this variable. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

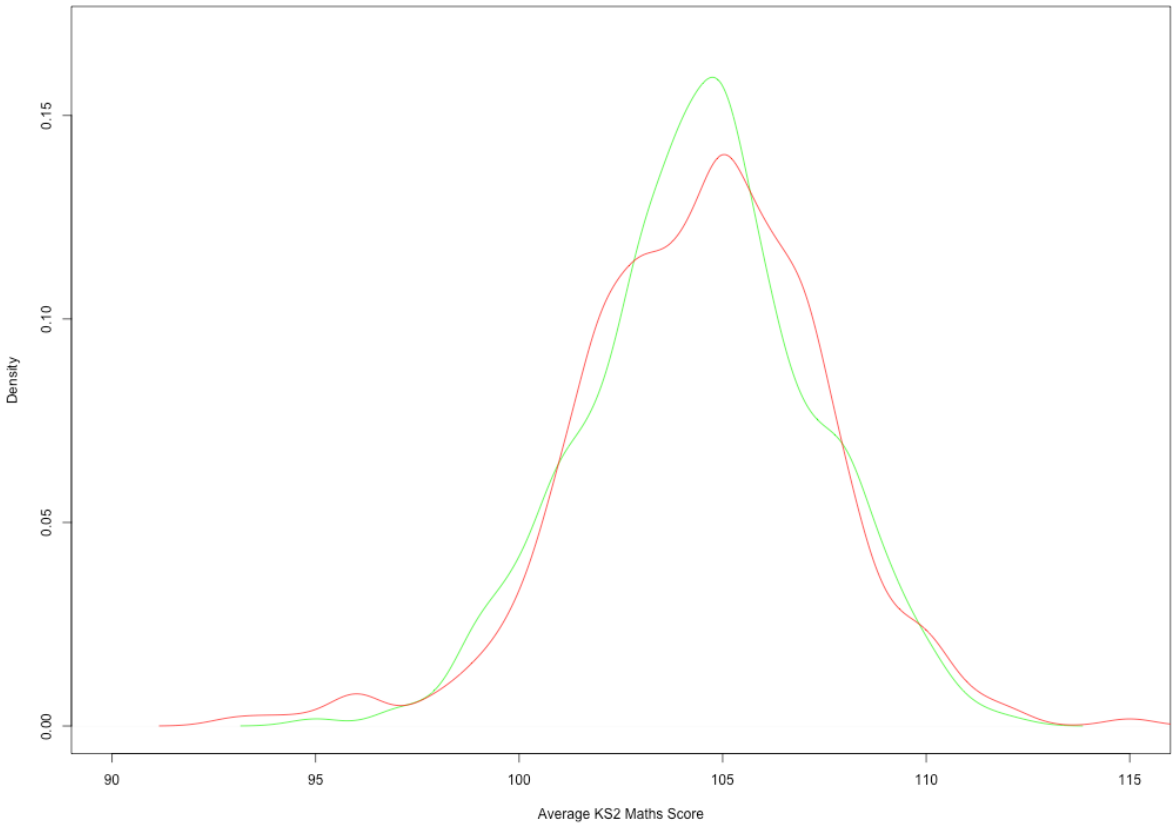
We demonstrate that the similarities in the means of continuous measures after matching is not hiding large differences in the distributions of the samples by plotting the full distribution of these variables in the treated and matched comparison samples. This is done for KS1 average points score of intake in Figure 1, for average KS2 maths score in Figure 2, for average KS2 reading score in Figure 3, for average KS2 maths score among FSM pupils in Figure 4, and for average KS2 reading score among FSM pupils in Figure 5. Unsurprisingly, given that they are no explicitly included in the matching model, the distributions are not quite as closely matched among FSM pupils but still perform acceptably.

Figure 1. Distribution of average KS1 points score of intake in treated and comparison groups



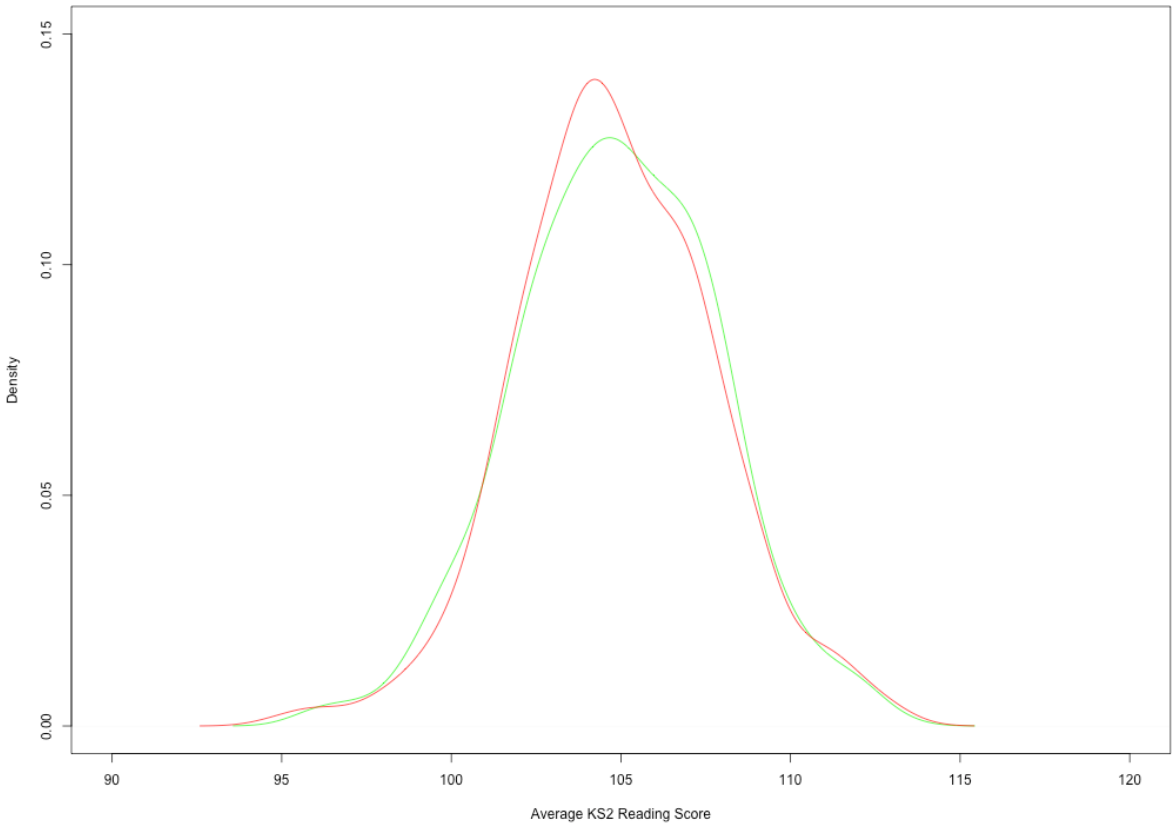
Notes. Kernel density plot of school average KS1 score of intake for treated (green) and comparison (red) schools.

Figure 2. Distribution of average maths score in treatment and comparison groups



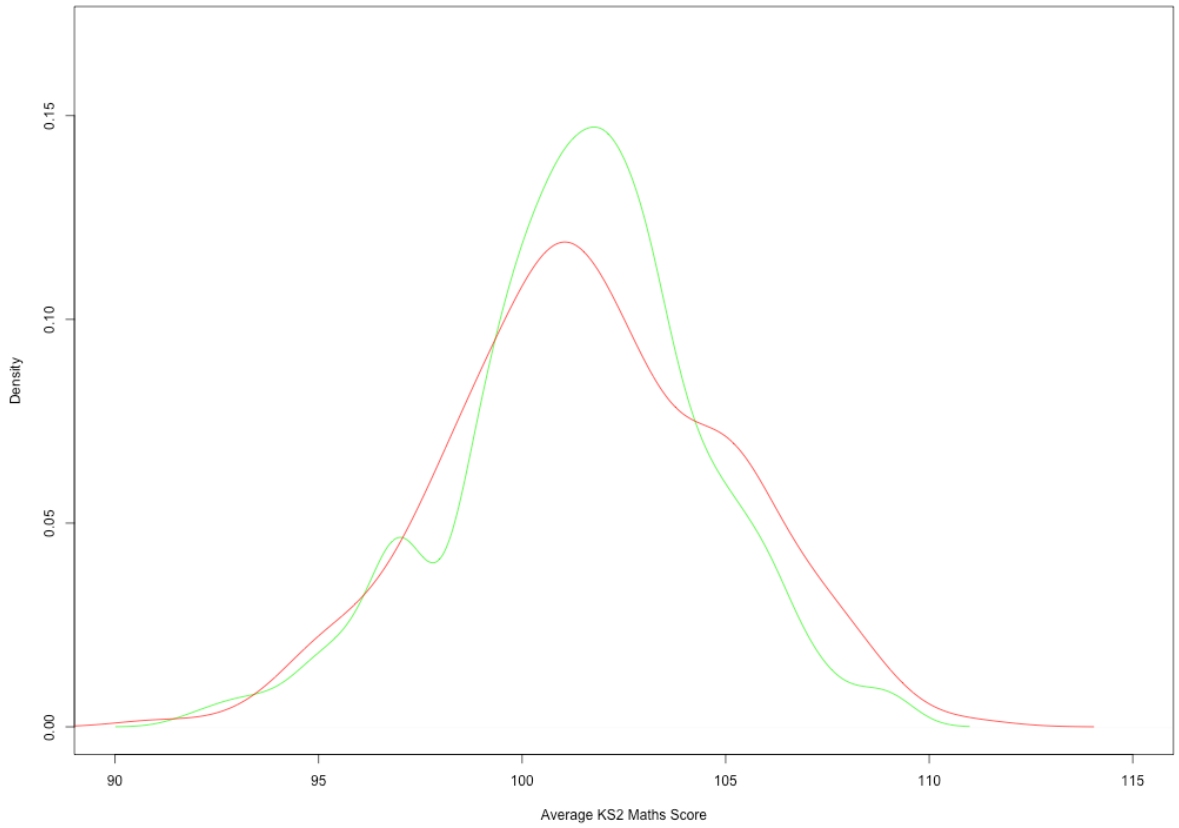
Notes. Kernel density plot of school average KS2 maths score for treated (green) and comparison (red) schools.

Figure 3. Distribution of average reading score in treatment and comparison group



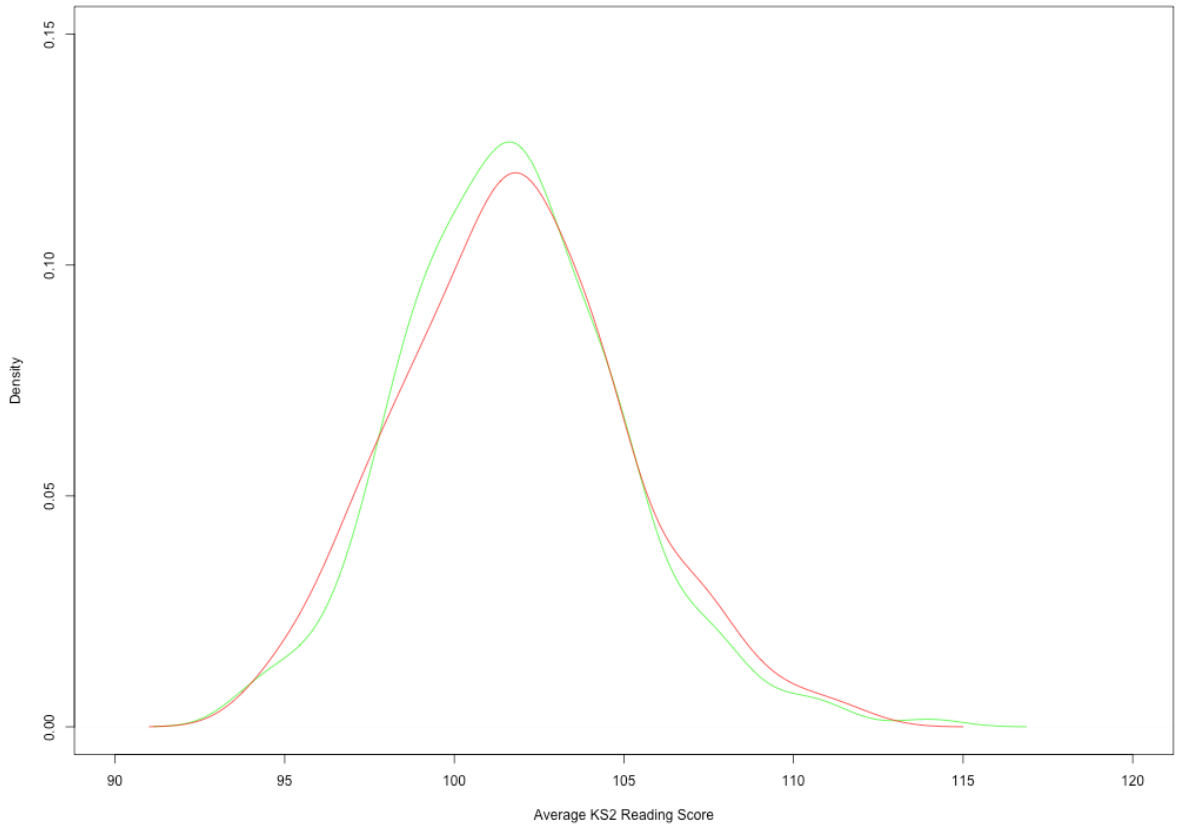
Notes. Kernel density plot of school average KS2 reading score for treated (green) and comparison (red) schools.

Figure 4. Distribution of average maths score among FSM pupils in treatment and comparison groups



Notes. Kernel density plot of school average KS2 maths score among FSM pupils for treated (green) and comparison (red) schools. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

Figure 5. Distribution of average reading score among FSM pupils in treatment and comparison group



Notes. Kernel density plot of school average KS2 reading score among FSM pupils for treated (green) and comparison (red) schools. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

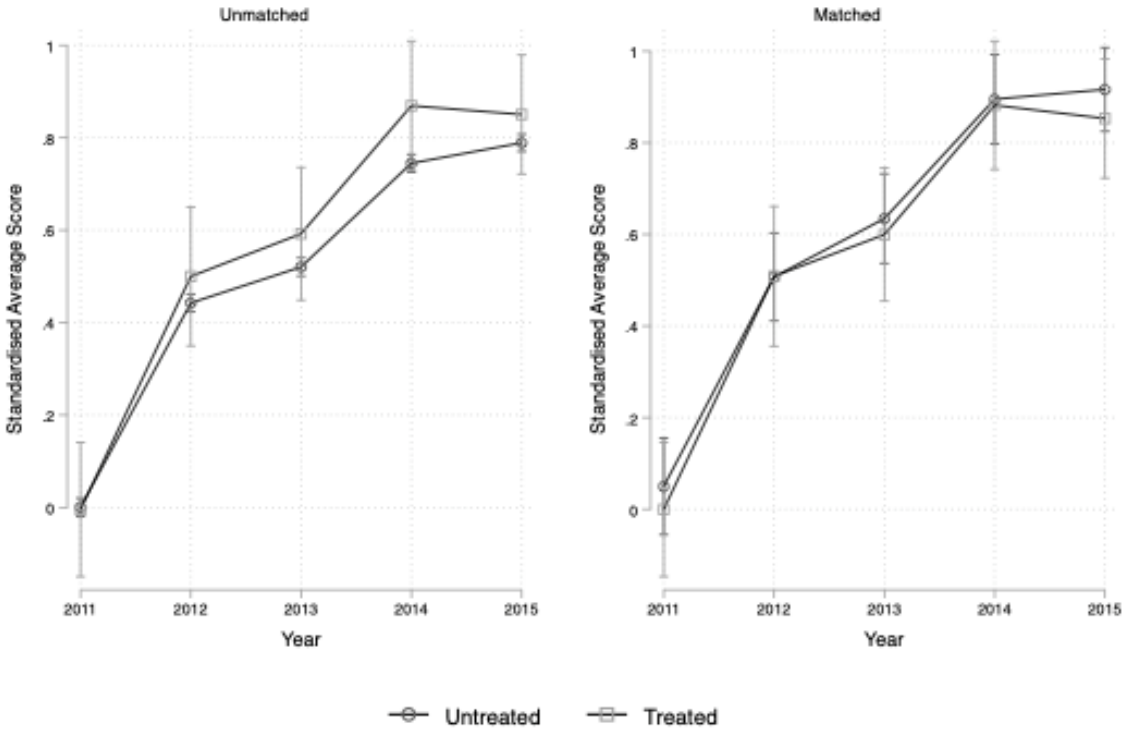
Common trends

In this section, we explore evidence of pre-treatment common trends in the performance of our treated schools, compared with the rest of schools considered for matching, and compared with our preferred matched sample. It should be remembered that this evidence is intended to explore the plausibility of the identifying assumption of difference in differences but does not and cannot "prove" this untestable assumption, which is that there would have been common trends between the two groups in the absence of our treatment. It should also be recalled that our matching exercise does not match directly on these trends but rather on observable school characteristics that we think are likely to have resulted in them being recruited into the study and, equivalently, on observable school characteristics that are likely to result in similar trends in their performance over time.

Figure 6 plots the KS2 average points score in treated and comparison schools before and after matching, while Figure 7 does this for KS1 points score of intake, Figure 8 for KS2 maths scores and Figure 9 for KS2 reading scores. Figure 10 and 11 plot these final two average KS2 maths and reading scores among FSM pupils only. Each shows first the trends before matching in the left hand panel; these plots demonstrate that, in fact, even before matching there are fairly similar trends between the treated schools and all others that could have been selected. This suggests that the developer team have done a good job of recruiting schools that are quite representative of the wider school population in the recruitment areas in terms of their performance trends. Each figure's right hand panel shows the trends after matching (the plotted line for treated schools barely changes since barely any treatment schools are discarded in the matching process). Here, the trends generally match one another even more closely, which we believe provides strong suggestive evidence in favour of our study's identifying assumption.

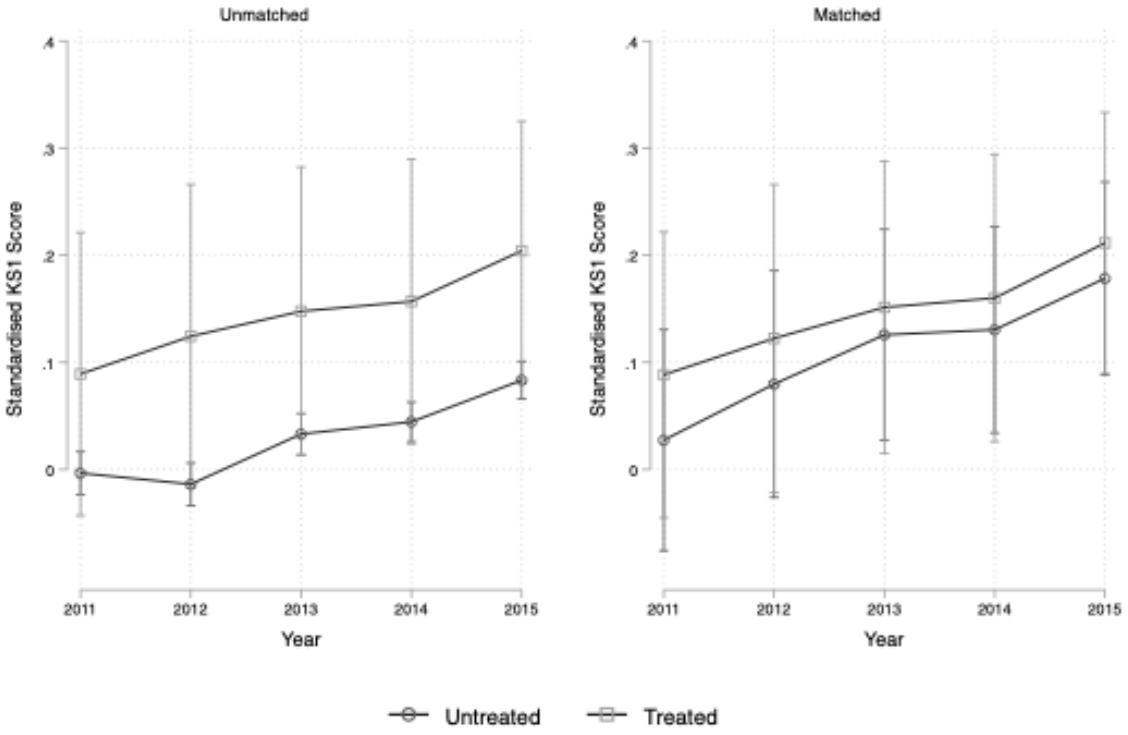
Given the similarity in common trends even in the unmatched sample, we propose to estimate the impact of the treatment using our difference in differences method using the unmatched sample in recruitment areas as an additional robustness check of our results.

Figure 6. Average points score 2011-2015 in treatment and comparison schools



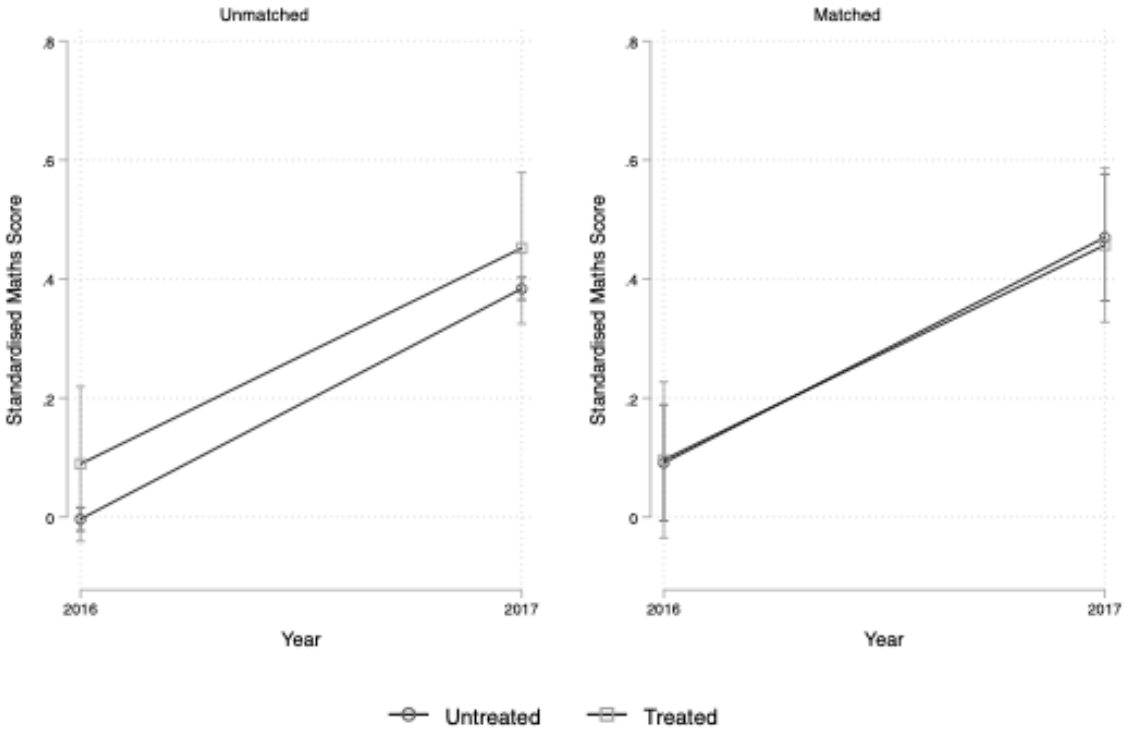
Notes. School KS2 average points score in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample.

Figure 7. Average KS1 points score of intake 2011-2015 in treatment and comparison schools



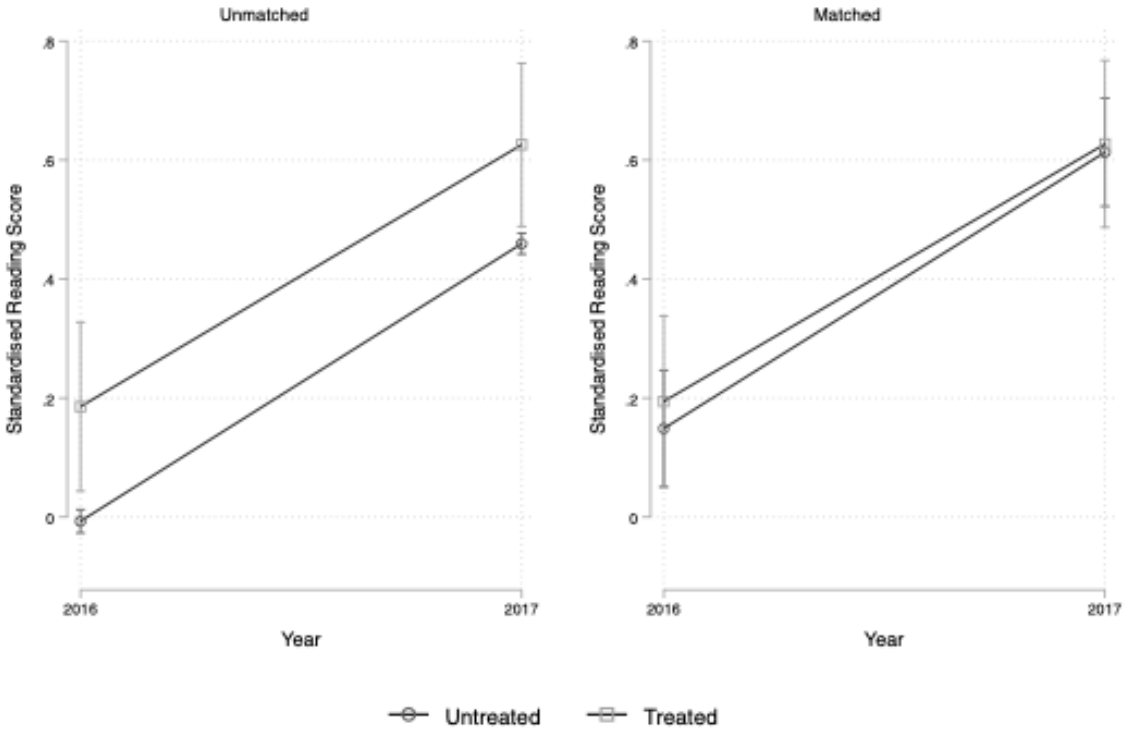
Notes. School KS1 average points score of intake in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample.

Figure 8. Average maths score 2016-2017 in treatment and comparison schools



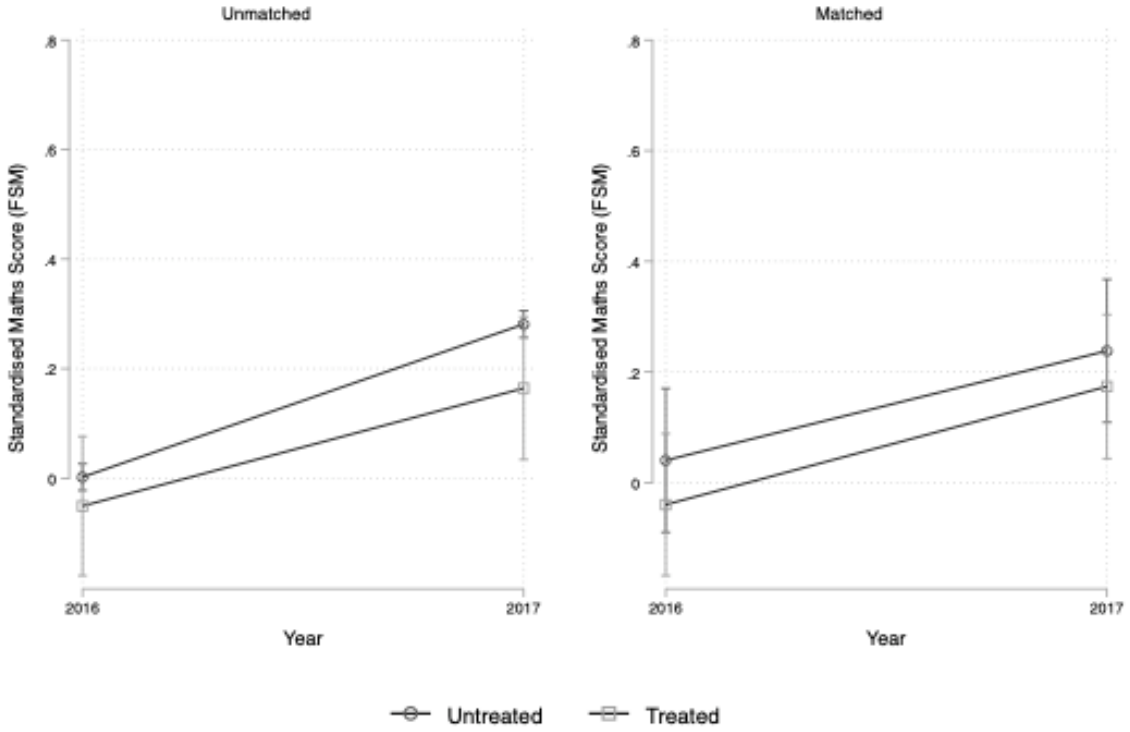
Notes. School KS2 average maths score in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample.

Figure 9. Average reading score 2016-2017 in treatment and comparison schools



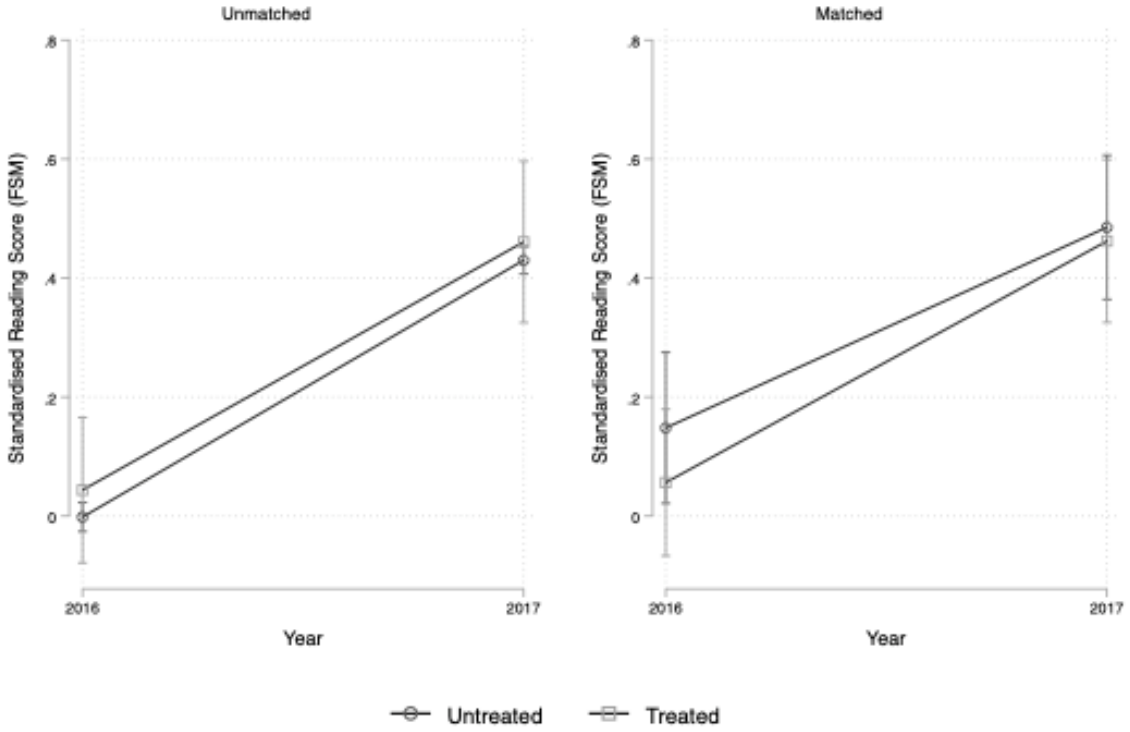
Notes. School KS2 average reading score in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample.

Figure 10. Average maths score for FSM pupils 2016-2017 in treatment and comparison schools



Notes. School KS2 average maths score among FSM pupils in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

Figure 11. Average reading score for FSM pupils 2016-2017 in treatment and comparison schools



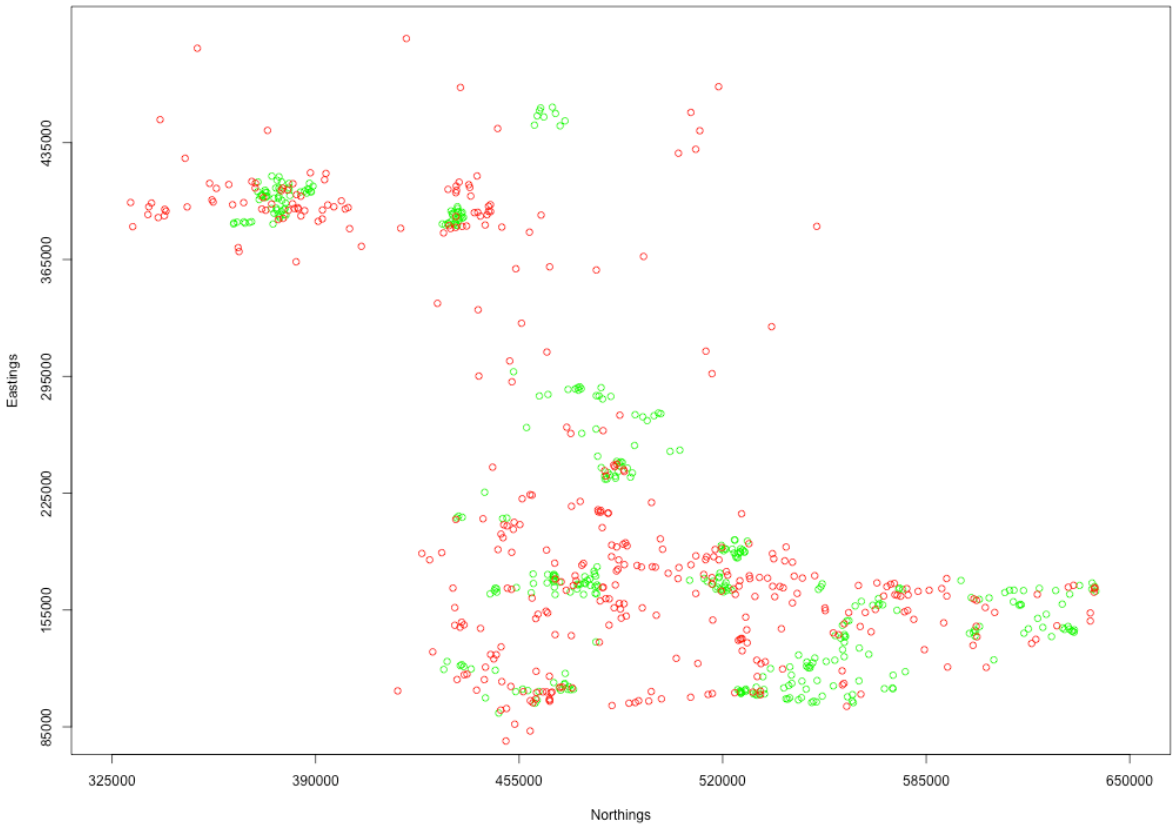
Notes. School KS2 average reading score among FSM pupils in treated and comparison schools with cluster-adjusted confidence intervals. Range of years reflects availability of consistent school-level measures. Scores have been standardised to have mean zero and standard deviation one in the first year of data availability in the overall sample. Note that FSM characteristics are estimated from a reduced sample size due to suppression in schools with small numbers of FSM pupils.

Geographical distribution of schools

Given the importance of geography in the formation of school clusters, we explored the importance of this factor in predicting involvement in this trial. Ultimately, in our preferred specification we use exact matching on region and urbanity/rurality of schools to maximise comparability in this regard. However, we also explored more direct modelling of location using a generalised additive model to specify the joint two-dimensional relationship between school location (i.e. longitude and latitude) and involvement in this research.

Figure 12 plots the location of the treatment and comparison schools in the preferred matched sample, as noted identified using matching on region and urbanity/rurality. It is evident that our matched sample still selects schools that are predominantly located in similar areas of England, although there are a number of outliers, which are possible despite the constraints imposed.

Figure 12. Location of treatment and comparison schools in matched sample identified using preferred specification

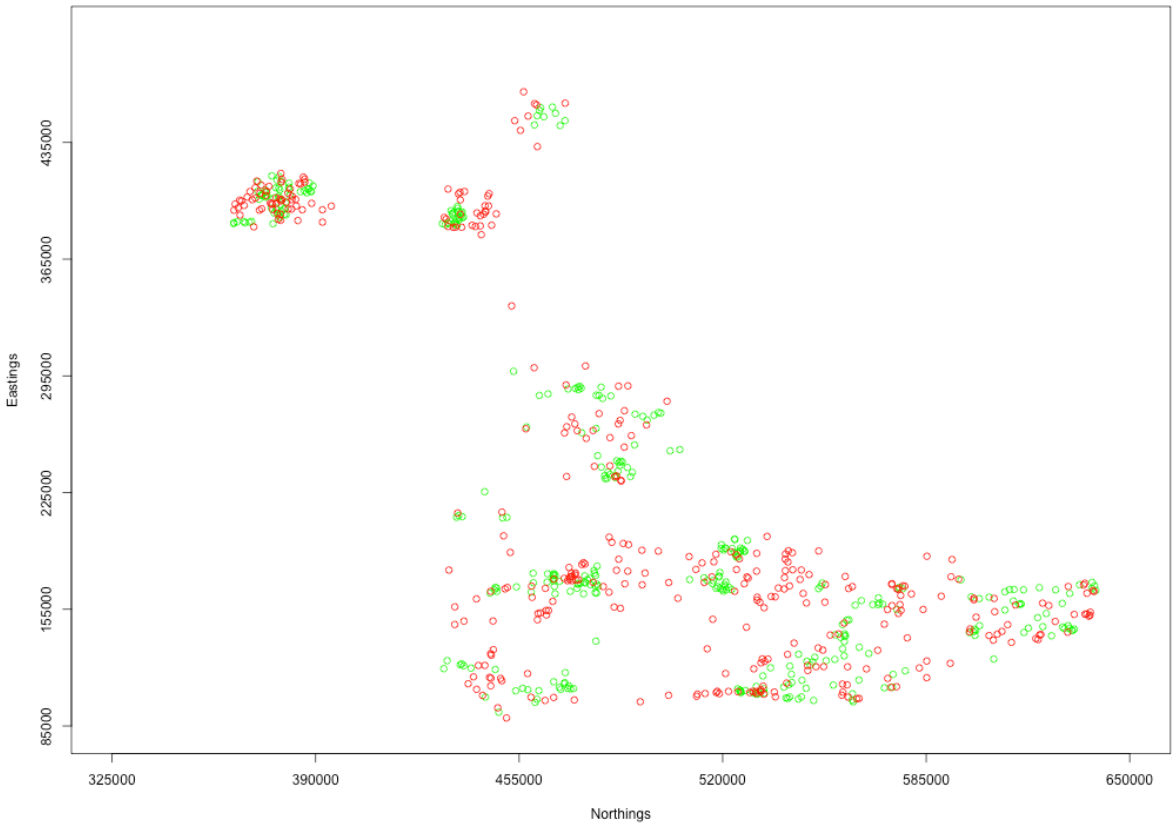


Notes. Geographical location of schools plotted using Nothings and Eastings grid references for treated (green) and comparison (red) schools.

Figure 13 plots the location based on our alternative approach to modelling location, which results in matched comparison schools located much closer to the treated schools. This approach clearly does

a much better job in this regard. However, while we expect geography to be important here, it is not the only important thing and it may well be the case that allowing matching to schools that are further away will be better matched in terms of other characteristics of important.

Figure 13. Location of treatment and comparison schools in matched sample identified using generalized additive modelling of school location



Notes. Geographical location of schools plotted using Nothings and Eastings grid references for treated (green) and comparison (red) schools.

Ultimately, given what we see as a broadly acceptable geographical distribution of matched comparison schools based on matching of region and urbanity/rurality, we maintain this as our favoured approach in order not to over-prioritise geography in our identification of an appropriate comparison group.

Appendix C. Logic Model