Statistical Analysis Plan
Schools Partnership Programme

Evaluator: UCL Institute of Education
Principal investigator: Jake Anders

Education Endowment Foundation

PROJECT TITLE	Schools Partnership Programme
DEVELOPER	Education Development Trust
EVALUATOR	UCL Institute of Education
PRINCIPAL INVESTIGATOR	Jake Anders
SAP AUTHORS	Jake Anders, John Jerrim
STUDY DESIGN	School-level matched difference in differences
STUDY TYPE	Effectiveness
AGE RANGE	10-11 (Key Stage 2)
NUMBER OF SCHOOLS	437
NUMBER OF PUPILS	N/A
PRIMARY OUTCOME MEASURE AND SOURCE	KS2 SATS maths performance from school-level NPD
SECONDARY OUTCOME MEASURE AND SOURCE	KS2 SATS reading performance from school-level NPD
EVALUATION PROTOCOL URL OR HYPERLINK	Link to evaluation protocol

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	18 March 2020	N/A

Table of contents

Introduction

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.

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).[1] 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).[2]

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).[3] 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).[4] This accords with the outcomes of an international comprehensive survey of assessment and evaluation in 28 countries by the OECD (2013).[5] 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.

The evaluation has been 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). Recruitment occurred in Winter/Spring 2017/18 with the aim of starting the intervention in September 2018. The evaluation will look at the school-level impact of the programme on KS2 maths tests (mat_average; primary outcome) and KS2 reading tests (read_average).

Design overview

Evaluation design	Matched difference in differences

[1] Burns, T. and Köster F. (eds.) (2016), Governing Education in a Complex World, Educational Research and Innovation, OECD Publishing, Paris.
[2] DfE (2010), The Importance of Teaching: The Schools White Paper, Cm 7980, Department for Education, London.
[3] Greany, T., and Higham, R., (2018) Hierarchy, Markets, Networks and Leadership: analysing the 'self-improving school-led system' agenda in England. IOE Press: London.
[4] 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.
[5] OECD (2013), Synergies for Better Learning: An International Perspective on Evaluation and Assessment. PISA, Paris: OECD.

Unit of analysis		School
Matching variables		Reading and maths average attainment in the school in 2017, average KS1 attainment of those who took KS2 tests in 2017, Number of pupils, Academy status, Most recent Ofsted inspection, IDACI quintile, Government Office Region
Primary outcome	variable	Maths attainment
	measure (instrument, scale)	Average Key Stage 2 Maths Scores (mat_average)
Secondary outcome(s)	variable(s)	Reading attainment
	measure(s) (instrument, scale)	Average Key Stage 2 Reading Scores (read_average)

This evaluation has been 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). This approach to the impact evaluation has been 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.

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 will receive the intervention, while statistical matching methods will be used to identify the counterfactual group. In order to be considered, schools had to agree to cooperate with the project and evaluation teams during the trial. The project team advertised the trial and also approached schools through their existing networks. Where possible they aimed to recruit schools that have larger populations of individuals receiving FSM.

In our preferred specification, matches are found using a nearest neighbour algorithm[6] 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 using the following characteristics: average KS2 reading score of the school in 2017, average KS2 maths score of the school in 2017, average KS1 score of the school's intake (among those taking KS2 tests in 2017), the number of pupils in the school in 2017, whether the school is an academy in 2017, school's most recent Ofsted rating in 2017, the quintile group into which the school falls in terms of the average Index of Deprivation Affecting Children and Infants (IDACI) of its intake, and the government office region in which a school is located.

[6] A nearest neighbour algorithm was chosen to prioritise the intuition and transparency of the selecting comparison schools.

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.[7]

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. These characteristics were chosen for exact matching to ensure we compare between schools with similar contexts and intakes, but without resulting in large reductions in the sample for whom we are able to find a match: 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).[8] Schools outside the range of common support were excluded, although when we tested removing this restriction this revealed that it was only resulting in one treatment school being excluded.

Follow-up

As reported above, 437 schools were recruited. However, missing administrative data on the matching variables meant that the sample considered for matching was 383 schools. In addition, a further 9 treatment schools were dropped as part of the matching process due to the imposition of common support and a caliper, resulting in a final analysis sample of treatment schools of 374; these will be compared to 374 matched comparison schools identified as described above. We regard both of these as essentially random processes marginally reducing the achieved sample (and, hence, statistical power the analysis will achieve) rather than being attrition from the trial.

Sample size calculations overview

		Protocol		Matching	
		OVERALL	FSM	OVERALL	FSM
MDES		0.20	0.20	0.17	0.17
Pre-test/ post-test correlations	level 1 (school)	0.16	0.16	0.16	0.16
	level 1 (cluster)	0.49	0.49	0.49	0.49
Intracluster correlations (ICCs)	level 1 (school)	N/A	N/A	N/A	N/A
	level 2 (cluster)	0.10	0.10	0.10	0.10
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided

[7] Daw, J. R., & Hatfield, L. A. (2018). Matching and Regression to the Mean in Difference-in-Differences Analysis. Health Service Research. doi:10.1111/1475-6773.12993
[8] 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

Average cluster size		6	6	4	4
Number of clusters	intervention	50	50	85	85
	comparison	50	50	85	85
	Total	100	100	170	170
Number of schools	intervention	300	300	374	374
	comparison	300	300	374	374
	Total	600	600	748	748

We conducted our protocol sample size calculation for the KS2 maths outcome, since this is the primary outcome. 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[9] and 6 schools within each cluster.

In conducting this calculation, using the formula reported in McConnell & Vera-Hernández (2015),[10] 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.[11]

These requirements suggest 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 we are using school-level variables, the power calculation for average performance of FSM pupils is no different to that for the overall outcome.[12]

These figures have been updated based on the recruited sample that were successfully matched in the preferred matching specification i.e. 374 schools in 85 clusters. Given the long tail in the distribution of cluster sizes (arithmetic mean = 5; median = 4; minimum = 2; maximum = 16) to be conservative, we have used the median of cluster sizes (also approximately equal to the harmonic mean) for the purposes of updating the power calculation. Other assumptions have been maintained as at protocol stage.

[9] 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.
[10] McConnell, B. and M. Vera-Hernández (2015). Going beyond simple sample size calculations: a practitioner's guide. IFS Working Paper Series. London, UK, Institute for Fiscal Studies.
[11] 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.
[12] 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

Primary outcome analysis

We will estimate the effect of the intervention using a linear model on school- and year-level data from the pre- and post-treatment periods (specifically academic year ending 2018 and academic year ending 2019). Raw outcome variables from the NPD, as described in the outcome measures section above, will be used in all models. Cluster-level clustered standard errors will be calculated in order to take into account the potential dependence of the results among school clusters; schools in the matched comparison group will be treated as independent from one another for the purposes of calculating standard errors.

The model will 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). The model will be specified as follows:

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

where $mat\_average$ is average KS2 maths score in school i in year t, $Treat$ is a binary variable indicating that schools are in the treatment rather than the matched comparison group, $Post$ is a categorical variable indicating the data comes from the post-treatment period i.e. 2020, $tks1average$ is the average KS1 intake score in school i in year t (i.e. for the cohort completing KS2 in this year), and ε_{it} is an idiosyncratic error term.

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

Secondary outcome analysis

We will conduct three secondary outcome analyses:

- **Reading performance:** Same as the primary outcome analysis except replace Y_{ij} with the KS2 average reading performance (school performance measures `read_average`).
- **Maths performance among FSM sub-group:** Same as the primary outcome analysis except replace Y_{ij} with the KS2 average reading performance (school performance measures `mat_average_fsm6cla1a`)
- **Reading performance among FSM sub-group:** Same as the primary outcome analysis except replace Y_{ij} with the KS2 average reading performance (school performance measures `read_average_fsm6cla1a`)

Interim analyses

No interim analyses are planned, other than the matching exercise itself and the imbalance checking from this, which was reported as Appendix B to the project's evaluation protocol, published as version 2 of this document on the EEF website in February 2019.

Subgroup analyses

No subgroup analyses are planned. As this is a school-level analysis, outcomes for the sub-group of FSM pupils are still carried out on the full sample of schools.

Additional analyses

As a check on the robustness of our findings from the preferred matching specification, the analyses reported above will be repeated using analysis samples that were constructed using the following alternative matching specifications. These samples are as follows:

- Caliper of 0.4: Sample of 374 treatment and 374 comparison schools
- 2 nearest neighbours: Sample of 374 treatment and 743 comparison schools
- No exact matching: Sample of 374 treatment and 374 comparison schools
- Exact matching also on Ofsted judgement: Sample of 366 treatment and 366 comparison schools
- GAM of school location replaces exact matching on urban/rural: Sample of 362 treatment and 362 comparison schools

Further details of these alternative specification and samples are reported in the matching exercise reported as Appendix B to the project's evaluation protocol, published as version 2 of this document on the EEF website in February 2019.

Imbalance at baseline

Imbalance at baseline will be reported using absolute standardised differences (Imbens & Rubin, 2015)[13] (i.e. the absolute value of the mean difference divided by the sample standard deviation)[14] between the treatment and comparison groups. These provide a simple, scale-free measure of differences that is easy to interpret. Analysis of the imbalance has already been carried out following the matching exercise and are reported in Appendix B of the updated evaluation protocol; they will be re-presented in the final report. Imbalance statistics cover the following school characteristics:

- KS2 Reading Score in 2017
- KS2 Maths Score in 2017
- KS2 Reading Score for FSM pupils in 2017
- KS2 Maths Score for FSM pupils in 2017
- KS1 Intake Score in 2017
- Academy status
- IDACI quintiles
- Ofsted rating

We will also demonstrate that the post-matching similarity in the means of the continuous measures above is not disguising large differences in the distributions of the samples by creating overlapping kernel density plots of the full distribution of these variables in the treated and matched comparison samples. These will be reporting in an appendix to the final report.

[13] Imbens, G. M. and D. B. Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY, Cambridge University Press.
[14] Standardised differences are practically the same as effect sizes but are conceptually different, since they are not attempting to quantify an effect.

Missing data

It is hoped that missing data will be minimal, since we are relying solely on publicly available administrative data available from the UK Department of Education. Nevertheless, we are aware of the possibility of missing data for a variety of reasons, including suppression of values in specific years due to small sample sizes and not being able to find longitudinal data about schools in our sample due to changes in identifier (although we will, of course, attempt to minimise these).

We will describe and summarise the extent of any missing data in the primary and secondary outcomes, and in the model associated with the analysis. Where possible, reasons for any missing data will also be described.

If more than 10% of data in the model is missing (based on the finalised matched sample i.e. 748 schools in the case of our primary analysis), we will implement the following missing data strategy. The strategy will be followed separately for each instance of model and variable for which the threshold is exceeded. We will first explore whether there is evidence that the missing data is missing at random (MAR), since this is a pre-requisite for missing data imputation modelling to produce meaningful results. To do this we will create an indicator variable for each variable in the impact model specifying whether the data is missing or not. We will then use logistic regression to test whether this missing status can be predicted from the variables used for imbalance testing (listed above). Where predictability is confirmed we will proceed to use these same variables to estimate a Multiple Imputation (MI) model using a fully conditional specification, implemented using Stata MI to create 20 imputed data sets; we believe this is an appropriate number of imputed datasets given the anticipated level of potential missing data as a result of the administrative data source we are employing.[15] We will re-estimate the treatment effect using each dataset and take the average and estimate standard error using Rubin's combination rules.[16]

Analysis using the dataset produced through either of these missing data strategies will be used as a sensitivity analysis i.e. we will base confirmation of the effectiveness of the treatment on complete case analysis only but assess the sensitivity of the estimate to missingness using the estimates from the multiply imputed dataset. If the complete case analysis model implies effectiveness but the imputed dataset analysis model does not (or changes the direction of the estimated effect) we must assume that the missing data is missing not at random to such an extent as to invalidate our conclusion of effectiveness, which we would state in the reporting of the evaluation.

Compliance

We will 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). Further details were provided in the evaluation protocol.

School-level categories and criteria

[15] Graham, J.W., Olchowski, A.E. & Gilreath, T.D. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prev Sci 8, 206–213 (2007). https://doi.org/10.1007/s11121-007-0070-9

[16] Rubin, D. (2004). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Categories	Attendance to training/workshops	Review visits
Non-compliant schools	No attendance/dropped out	No review visits/dropped out
Minimally compliant schools	Attendance to less than 75% of training sessions/workshops	Has hosted 1 peer review visit across programme
Fully compliant schools	Attendance to 75% or more of training sessions/workshops	Has hosted 2 review visits across programme
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 categories and criteria

Cluster-level compliance categories are based on aggregation of school-level compliance of all schools within a cluster. 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**.

Analysis

It is unclear the properties of using a method such as instrumental variables to conduct compliance analysis within the scope of our evaluation design. As such, we instead stick to repeating our primary analysis among the following sub-groups:

- All treatment schools judged to be at least minimally compliant and school-to-school matched comparators
- All treatment schools judged to be fully compliant and school-to-school matched comparators
- All treatment schools in clusters that are judged to be at least minimally compliant and school-to-school matched comparators
- All treatment schools in clusters that are judged to be fully compliant and school-to-school matched comparators

Nearest neighbour matching is not designed to identify which treatment school is matched with which comparison school. As such, we turn to optimal matching (on the propensity score estimated for the main matching exercise) within the analysis sample to identify school-to-school matched comparators for this purpose. The code that will be used for this purpose is included in Appendix B.

Intra-cluster correlations (ICCs)

In order to estimate the intra-cluster correlation (ICC) of the outcome measures at cluster-level we will employ an empty variance components model, as follows:

$$\boldsymbol{Y_{ij} = \alpha + \eta_j + \varepsilon_{ij}}$$

where school $\boldsymbol{i}$ is nested in cluster $\boldsymbol{j}$, $\boldsymbol{Y_{ij}}$ is the average KS2 maths score for the purpose of calculating the post-test ICC, $\boldsymbol{\eta_j}$ is a cluster-level random effect, and $\boldsymbol{\nu_{ij}}$ is an school-level error term. The cluster-level random effect is assumed to be normally distributed and uncorrelated with the school-level errors.

The ICC itself will be estimated from this model using the following equation:

$$\boldsymbol{\rho = \frac{var(\eta_j)}{var(\eta_j)+var(\varepsilon_{ij})}}$$

Effect size calculation

Hedges' g effect size will be calculated as follows:

$$g \ = J(n_1 + n_2 + 2)\frac{\overline{x_1} - \overline{x_2}}{\widehat{s*}}$$

where our conditional estimate of $\overline{x_1} - \overline{x_2}$ is recovered from $\beta_1$in the primary ITT analysis model;

$\widehat{s*}$ is estimated from the analysis sample as follows:

$$s^* \ = \sqrt{\frac{(n_1 - 1){s_1}^2 \ + \ (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation of the control group, and s_2 is the standard deviation of the treatment group (all estimates of standard deviation used are unconditional, in line with the EEF's analysis guidance to maximise comparability with other trials);

and $J(n_1 + n_2 + 2)$ is calculated as follows:

$$J(n_1 + n_2 + 2) = \frac{\Gamma\left(\frac{n_1 + n_2 + 2}{2}\right)}{\sqrt{\frac{n_1 + n_2 + 2}{2}}\ \Gamma\left(\frac{n_1 + n_2 + 2 - 1}{2}\right)}$$

where n_1 is the sample size in the control group and n_2 is the sample size in the treatment group.

If calculating $J(n_1 + n_2 + 2)$ proves computationally intractable[17] using the above method, we will instead use the following approximation:

$$J(n_1 + n_2 + 2) \approx (1 - \frac{3}{4(n_1 + n_2) - 9})$$

Ninety-five per cent confidence intervals (95% CIs) of the effect size will be estimated by inputting the upper and lower confidence limits of $\widehat{\beta_1}$ from the regression model into the effect size formula.

All of these parameters will be made available in the report.

[17] The output of the gamma (Γ) function in the Hedges' g correction factor (J) becomes large quickly, making this method of computation intractable where $n_1 + n_2$ is not small. As such, it can quickly become intractable. Thankfully, the approximate method tends towards the fully correction factor quickly. As such, where the computational intractability is an issue the approximate method is appropriate. In any event, the correction factor is likely to be small in this trial.

Appendix A: Analysis Syntax

In this appendix, we provide indicative analysis syntax to implement the models specified in the Statistical Analysis Plan using Stata 16. Eventual syntax may have small changes (e.g. variable name changes) that do not affect the syntax's implementation of the models specified above. Variables are as specified in the statistical analysis plan.

All analyses will use variants of the following linear regression model, albeit estimated on different samples or using different outcome variables:

```
regress mat_average treat post treat#post tks1average, vce(cluster
cluster_id)
```

estimated on school- and year-level matched sample data pooled across 2018 and 2020 where `mat_average` is the school's average KS2 maths score (corresponding to $mat\_average$ in the regression equation) in a given year, `treat` is a binary variable indicating that schools are in the treatment rather than the matched comparison group (corresponding to $Treat$ in the regression equation), `post` is a categorical variable indicating the data comes from the post-treatment period i.e. 2020 (corresponding to $Post$ in the regression equation), `tks1average` is the school's average KS1 intake score for the relevant cohort (corresponding to $tks1average$ in the regression equation), and `cluster_id` is a cluster identifier.

Appendix B: School-to-school matching code

```
matchforcompliance <- matchit(treated ~ pscore + 1,
                         method = "optimal", discard="none",
                         ratio=1,
                         data=input)
sample <- match.data(matchforcompliance)
```