Statistical Analysis Plan for The RISE project

UCL Institute of Education

INTERVENTION	**The RISE Project: Evidence-informed school improvement**
DEVELOPER	Huntington school
EVALUATOR	UCL Institute of Education
TRIAL REGISTRATION NUMBER	ISRCTN 18127836
TRIAL STATISTICIAN	John Jerrim
TRIAL CHIEF INVESTIGATOR	Meg Wiggins
SAP AUTHOR	John Jerrim
SAP VERSION	1
SAP VERSION DATE	27/10/2017

Protocol changes

1. In the study protocol, there was a sole primary outcome, which was to combine the GCSE maths and English language scores into a single variable. This has been changed, with English and mathematics analysed separately, as dual primary outcomes.

Introduction

This project, led by Huntington School, aims to test whether a research-based school improvement model makes a significant difference to classroom practice and student outcomes. Each school in the programme will appoint a 'research lead' who will be responsible for implementing the improvement programme in their school, with a particular focus upon improving student attainment in English and Maths at GCSE. The research leads will be supported by a thorough programme of workshops delivered by the team from Huntington School, alongside a collaborative network-based approach to support. The Centre for Evaluation and Monitoring (CEM) at Durham University will support Huntington to develop and deliver the content of these workshops, and the guidance on designing appropriate, robust, school–led evaluations.

Study design

The evaluation is a stratified, clustered randomised controlled trail with two arms. Four strata of ten schools were created prior to randomisation based upon historic GCSE performance. Within each of the strata, five schools were assigned to treatment and five to the control group. Therefore, of the 40 participating schools, 20 receive the treatment and 20 the control.

There are two 'measurement points'. The first is pupils' Key Stage 2 test scores, reflecting their achievement as they finished primary school (these tests were therefore conducted three years before the intervention began in one-year intervention cohort and four years before for the two-year intervention cohort.). The motivation for using Key Stage 2 as a pre-test is that the data is free, and requires no burden upon schools. There is also little evidence that conducting an additional baseline test would significantly improve statistical power.

The second measurement point refers to children's grades in GCSE mathematics and English, which is directly after the intervention has finished for the groups.

Randomisation

The 40 participating schools were first divided into four strata, each containing ten schools. These strata were formed based upon a 3-year average of historical school level GCSE grades. Within each of these strata, five schools were randomly allocated to treatment and five to control. The randomisation was performed by John Jerrim, using a random number generator in Excel. In total, there are 20 treatment schools and 20 control schools.

As the baseline test scores in this trial are Key Stage 2 results, technically randomisation was conducted after the pre-test measure. However, as pupils and teachers would not have known they were to be part of this RCT when they took their Key Stage 2 examinations, this is unlikely to have an impact upon the results.

Calculation of sample size

We have estimated power calculations for outcomes relating to a) pupil attainment at GCSE and for b) teachers research use and understanding (as measured by the survey). These have been calculated using the optimal design software.

a. The intervention providers only have the capacity to provide the intervention to 20 schools. Agreement was made that the total sample size of schools would therefore be limited to 40. For pupil attainment in English and maths at GCSE, with school level randomisation of 40 schools, we have assumed an intra-cluster (between pupils within schools) correlation of 0.15, 200 pupils per school year group, a baseline of KS2 SATS, and a pre-post correlation (KS2 to KS4) of 0.7 (authors' calculations using NPD database). Given these assumptions, if 40 schools were retained in the trial, one could detect a minimum effect size of approximately 0.355. This is a large MDES, and we believe the RCT to be substantially underpowered.

However, as the intervention providers only had capacity to deliver the intervention in 20 schools, little could be done to overcome this challenge.

b. For teacher research understanding and use outcomes using the NfER devised research use tool, with school level randomisation of 40 schools, we have assumed an intra-cluster (between schools) correlation of 0.05, 20 teachers per school, a baseline using the same tool, and a pre-post correlation of 0.8. Given these assumptions, if 40 schools were retained in the trial, one could detect a minimum effect size of approximately 0.23 (80% power for 95% CU).

Follow-up

The baseline data collection exercise using the NfER tool with teachers and senior leadership team had a lower than desired response rate – 44% (427 of 979 returned). This suggests that follow up response will have issues with missing data. Two control schools have had limited engagement with the developer team and it is especially doubtful whether there will be response from them at follow up. To increase response rate at follow up, there will be multiple reminders of the on-line questionnaire, as well as a hard copy version of the questionnaire sent to continued non-responders. We will liaise with department leads to ask them to encourage their staff to complete the questionnaire.

Outcome measures

Primary outcome

The dual primary outcomes will be GCSE mathematics and GCSE English grades. One challenge with this trial is that is straddles across two year groups (2015/2016 and 2016/17 academic years) when the GCSE examinations are changing. The analysis will therefore be conducted separately by year group (see below for further details).

For pupils who took their GCSEs in the 2015/16 academic year, we will use the point score in the mathematics EBacc pillar (variable name KS4_EBPTSMAT_PTQ_EE from the NPD tables downloaded from https://www.gov.uk/government/publications/national-pupil-database-user-guide-and-supporting-information) and the Point score in English EBacc pillar (variable name KS4_EBPTSENG_PTQ_EE).

For the pupils who are taking their GCSEs in 2016/2017, we will use the analogous variables to those described in the paragraph above. However, given the changes due to be made to GCSEs, the precise variable name in the NPD is not yet known. This is also based upon the assumption that an analogous 'pillar score' for GCSE mathematics and English will be made available.

Note that the above are not truly continuous variables (they have only a limited number of categories – 10). In our analysis, we will therefore test the robustness of results to using an ordered logistic regression as an alternative to an OLS regression (See Analysis Section).

Secondary outcomes

The secondary outcome is teachers' knowledge based upon the research use outcomes survey tool devised by NFER (see Appendix A). This survey will be administered to teachers at study schools at two time points: baseline (pre-randomisation Autumn 2014) and post intervention (Autumn 2017). The sample will include all teachers from English and maths departments, as well as school senior leadership teams. Following NfER guidance on the analysis of the tool, the analysis will focus upon combined scores for six specific constructs[1], including:

1. Positive disposition to academic research in informing teaching practice (combining answers to items in questions 24 [item 3], 26 [1,3,5] & 27 [2,4] on survey - see Appendix A);

[1] Poet H, Mehta P, Nelson J. (2015) *Research Use in Schools: Survey, analysis and guidance for evaluators*. Slough: NFER.

2. Use of academic research to inform selection of teaching approaches (combining items in questions 23 [3] and, depending on filter for role, score on either question 6 [English], 11[maths], 17 [SLT] AND item 6 on either question 8 [English], 13 [maths] or 19 [SLT]
3. Perception that academic research is not useful to teaching (combining items in questions 26 [2] and 27 [3]);
4. Perception that own school does not encourage use of academic research (combining items in questions 26 [4] and 27 [1];
5. Active engagement with online evidence platforms (combining items in questions 23 [6] & 24 [6]);
6. Research knowledge (combining summed scores for all items questions A and B in Appendix A)

Each construct will sum the answer scores from the included items to form a raw score.

Analysis

The analysis will be conducted by John Jerrim who performed the randomisation. It will therefore not be conducted blind to group identity.

Primary intention-to-treat (ITT) analysis

Primary outcome

Within each participating school, there are two year groups who have received the intervention. The first year group will have received the intervention for one year only; the second year group will have received the intervention for two years. Given the challenge posed by the changes made to the GCSE outcome measure between the 2015/16 and 2016/17 cohorts, *all analyses will be conducted separately by year group.*

Our primary analysis model will take the form of the following OLS regression model:

$$Y_{Ij} = \alpha + \beta . T_j + \gamma . P_{ij} + \varepsilon_{ij} \qquad \nabla \mathrm{K}$$

Where:

Y_{Ij} = GCSE pillar points score (see above for details)

T_j = A school level dummy variable for treatment status (0 = Control; 1 = Treatment)

P_{ij} = The student level pre-test score (as measured by Key Stage 2 scores in English[2] and mathematics[3])

ε_{ij} = The error term. Clustering of pupils within schools will be accounted for by a Huber-White adjustment to the estimated standard errors. Stratification will also be accounted for by an adjustment made to the estimated standard errors.

I = Pupil i

J = School j

$\nabla \mathrm{K}$ = Indicating that the model will be estimated separately for the 2015/16 and 2016/17 cohorts.

Note that the use of stratification and clustering in the sample design will be accounted for via adjustments to the estimated standard errors. This will be done via the 'PSU' and 'strata' commands within Stata's 'svy' module. Stata will be used to conduct all the analyses. Following EEF guidance, we

[2] Variable KS4_VAP2TAENG_PTQ_EE in the NPD datafile
[3] Variable KS4_VAP2TAMAT_PTQ_EE in the NPD datafile.

will also supplement the above by estimating the treatment effect in terms of the sample difference in mean scores between treatment and control group.

For consistency with the vast majority of EEF trials, in our primary analysis we will treat the outcome measure (GCSE pillar scores) as a continuous variable. We will therefore be able to report results as per other EEF trials (e.g. effect sizes in terms of standard deviation units). However, as these variables are in reality only quasi-continuous (i.e. around 8 to 10 ordered categories) we will also estimate ordered logistic regression models to test the sensitivity of the results.

Analysis on both teacher and student outcomes will be by intention to treat (ITT). If contamination is present, a contamination-adjusted ITT (an instrumental variable approach) will supplement the main analysis.

Effect size calculations will be based upon the regression model specified above (i.e. 'adjusted differences' will be used) will be reported using Cohen's d and estimated separately for each regression model.

Interim analyses

Although sequential analysis may be possible in this trial (e.g. analysis could be conducted on the 2015/2016 cohort before data has arrived for the 2016/2017 cohort) this is not currently planned. Rather, all analysis will be conducted only once the 2015/16 and 2016/17 cohorts have completed their GCSEs.

Imbalance at baseline

For the primary analysis, 'balance' will be determined based upon Key Stage 2 mathematics and English scores, gender and FSM status as measured at the pupil level. (See footnotes 2 and 3 for details on exact Key Stage 2 variables that will be used). A 'threshold' equivalent to an effect size of 0.05 will be used as the threshold to determine whether 'balance' has been achieved upon these key pre-treatment characteristics. We will test the robustness of our results to including these as additional covariates in a supplementary model.

Missing data

As the primary outcome and baseline is based upon the NPD, we do not expect significant amounts of missing data. However, if more than 10% of baseline data are missing, we will investigate the patterns of missingness by cross-tabulating a missing indicator against the baseline pupil and teacher characteristics described above (Key Stage 2 English and Mathematics scores, FSM and gender).

Compliance analysis

Complier Average Causal Effect (CACE)[4] analysis will be used to explore dosage effect. It should be noted that this methodology essentially takes the ITT estimate and scales the estimated effect size upwards by the amount of non-compliance. Compliance measures will be included here once an agreement is reached with the developer.

Secondary outcome analyses

<u>Secondary outcome</u>

Our secondary outcome analysis will take the form of the following OLS regression model:

$$Y_{Ij} = \alpha + \beta . T_j + \gamma . P_{ij} + \varepsilon_{ij}$$

Where:

[4] Gerber AS, Green DP. (2012) Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

Y_{Ij} = Teachers' scores on each construct of the NFER research use survey

T_j = A dummy variable for treatment status (0 = Control; 1 = Treatment)

P_{ij} = Teachers' pre-intervention score on the NFER research use survey[5]

ε_{ij} = The error term. Clustering of pupils within schools will be accounted for by a Huber-White adjustment to the estimated standard errors. Stratification will also be accounted for by an adjustment made to the estimated standard errors.

I = Teacher i

J = School j

Note that the use of stratification and clustering in the sample design will be accounted for via adjustments to the estimated standard errors. This will be done via the 'PSU' and 'strata' commands within Stata's 'svy' module. Stata will be used to conduct all the analyses.

For the secondary outcome, we will investigate 'balance' between the treatment and control group in terms of teachers' baseline scores on the teacher research use survey.

In terms of 'sub-group' analysis, we will investigate possible heterogeneity in the impact of the intervention by teachers prior knowledge. This will be captured by an interaction between the treatment indicator (T) and teachers' pre-test scores on the NFER teacher knowledge measure.

For the secondary outcome, missing data could become a significant problem. If missing data is above 10%, we will investigate the patterns of missingness by cross-tabulating a missing indicator against the baseline pupil and teacher characteristics described above (Key Stage 2 English and Mathematics scores, FSM, gender, NFER teacher knowledge score). If there is missing data on the baseline measure (but not on the outcome) we will perform sensitivity analyses using multiple imputation via chained equations. Specifically, if more than 10% of baseline data is missing (but outcome data are available) we will perform both a complete case analysis and a multiple imputation analysis (using 10 imputed datasets). We do not feel it is advisable to specify the precise imputation to be used at this point, as this will in part be determined by the nature of the missingness (which is unknown apriori).

Additional analyses

None at present.

Subgroup analyses

For the primary outcome, we will investigate heterogeneous treatment effects by FSM (variable ever_fsm_6). This will be done by estimating a separate regression model for FSM pupils.

Effect size calculation

Effect sizes will be calculated via Cohen's d. This will be done by first converting the outcome into a z-scores (subtracting the sample mean and dividing by the sample standard deviation) and using this standardised variable in the analysis. Confidence intervals will be calculated after stratification and clustering has been taken into account via adjustments to the estimated standard errors (e.g. Huber-White to account for clustering).

As noted above, sensitivity analysis will be conducted re-estimating our analysis models using ordinal logistic regression. Effect sizes from these sensitivity analyses will be reported in terms of odds ratios.

[5] Will not be used in final construct: Research Knowledge, as this question is not asked at baseline.

Appendix A

Additional questions for teachers *outcomes* survey only:

Your knowledge about research

In this section we would like to gather some information about your knowledge of research. Please answer the questions without referring to other sources.

A. Current understanding from academic research suggests that each of the following statements is 'true' or 'false'. *(Please tick the answer that you know to be correct in each row. If you are not sure, please tick 'don't know').*

The research says that:	True 1	False 2	Don't know 3
Drinking six to eight glasses of water per day improves pupil learning outcomes	□	□	□
Reducing class size is one of the most cost-effective ways to improve pupil learning outcomes	□	□	□
Extending the school day is more likely to improve learning outcomes for pupils on Free School Meals than pupils not on Free School Meals	□	□	□
Interventions that focus solely on raising pupil aspirations have little impact on learning outcomes	□	□	□
Setting pupils by ability improves learning outcomes for all pupils	□	□	□
Individual pupils learn best when they receive information in their preferred learning style (e.g. auditory, visual, kinaesthetic)	□	□	□
Peer tutoring (students supporting other students with their learning) usually benefits the pupil being tutored more than the pupil doing the tutoring	□	□	□
Homework has a greater impact on pupils' learning outcomes at secondary school than at primary school	□	□	□

B. Below are descriptions of three reasons why someone would want to carry out research. Along the top of the table are five different research methods.

Please match the research purpose with the best research method for achieving it by selecting the relevant option. Please select one box in each row. There are only three matches – two methods are incorrect (please do not use the same answer more than once).

	Randomised Controlled Trial	Longitudinal study	Interviews and/or questionnaires	Literature review	Correlational study
To provide an **overview** of the evidence base	□	□	□	□	□
To determine **whether** an intervention or approach has a direct impact on pupil learning outcomes	□	□	□	□	□
To understand **how** an intervention or approach works in practice	□	□	□	□	□

Appendix B

Additional details on power calculations

Model 3.1: MDES Calculator for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2		
Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.15	Proportion of variance in outcome that is between clusters
P	0.50	Proportion of Level 2 units randomized to treatment: J_T / (J_T + J_C)
R_1^2	0.50	Proportion of variance in Level 1 outcomes explained by Level 1 covariates
R_2^2	0.00	Proportion of variance in Level 2 outcome explained by Level 2 covariates
g*	0	Number of Level 2 covariates
n (Average Cluster Size)	200	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
J (Sample Size [# of Clusters])	40	Number of Level 2 units
M (Multiplier)	2.88	Computed from T_1 and T_2
T_1 (Precision)	2.02	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.355**	Minimum Detectable Effect Size

Model 3.1: MDES Calculator for Two-Level Cluster Random Assignment Design (CRA2_2)— Treatment at Level 2

Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.15	Proportion of variance in outcome that is between clusters
P	0.50	Proportion of Level 2 units randomized to treatment: J_T / (J_T + J_C)
R_1^2	0.49	Proportion of variance in Level 1 outcomes explained by Level 1 covariates
R_2^2	0.25	Proportion of variance in Level 2 outcome explained by Level 2 covariates
g*	0	Number of Level 2 covariates
n (Average Cluster Size)	200	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
J (Sample Size [# of Clusters])	40	Number of Level 2 units
M (Multiplier)	2.88	Computed from T_1 and T_2
T_1 (Precision)	2.02	Determined from alpha level, given two-tailed or one-tailed test
T_2 (Power)	0.85	Determined from given power level
MDES	**0.308**	Minimum Detectable Effect Size