Statistical Analysis Plan
Same Day Intervention

Evaluator: The National Centre for Social Research
Principal investigator(s): Daniel Phillips

Template last updated: March 2018

PROJECT TITLE	Same Day Intervention
DEVELOPER (INSTITUTION)	The Yorkshire and Humber Maths Hub/Outwood (The Maths Hub is led by Outwood trained teachers, headteachers and teaching assistants)
EVALUATOR (INSTITUTION)	The National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR(S)	Daniel Phillips
TRIAL (CHIEF) STATISTICIAN	Robert Wishart
SAP AUTHOR(S)	Daniel Phillips, Robert Wishart, Anysia Nguyen
TRIAL REGISTRATION NUMBER	ISRCTN43822826
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/same-day-intervention/

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]	22/05/2019	*[leave blank for the original version]*

Table of contents

Contents

Introduction

This analysis plan sets out the detail of the analysis planned for the cluster-randomised controlled efficacy trial of Same Day Intervention (SDI).

The Same Day Intervention entails teachers and teaching assistants (TAs) receiving training in Same Day pedagogy, observing 'open classroom' sessions and receiving other support and access to teaching resources.

For the intervention itself, Same Day classes replace traditional mathematics classes. Each SDI class focuses on a single maths topic, with the principle goal being to ensure that by the end of the class, all pupils have a core understanding of the topic. Teachers demonstrate a topic, before pupils are given five or six questions to complete independently. There is then a 15 minute 'pit stop', during which teachers mark pupils' work, and pupils either attend a short assembly or are taught by a TA. After the break, pupils are grouped according to their diagnostic activity performance and there is an intervention session designed to target pupils who need extra teaching, address common misconceptions and embed learning. Same Day classes last 75 minutes, including the 15 minutes 'progress pit-stop'.

SDI aims to ensure all pupils have grasped the key elements of a topic by the end of a class. The 'progress pit-stop' facilitates the identification of pupils with misconceptions regarding maths concepts and allows teachers to address them and thereby reduce the learning gap. By ensuring children are not left behind, SDI aims to increase pupils' confidence in their maths ability. The targeted teaching is also intended to improve maths attainment for pupils. Finally, by incorporating marking into classes and reducing the number of pupils falling behind (and therefore needing additional support), it is hoped that SDI may reduce teacher workload.

The evaluation will be conducted as a two-arm cluster (school-level) randomised controlled efficacy trial. The primary outcome of interest is maths attainment as measured by GL's Progress Test in Maths and secondary outcomes are teacher perceptions of pupils' maths confidence and teacher workload, both measured by a survey.

Specifically, the trial aims to answer the following research questions:

- What is the impact of SDI on maths attainment of Year 5 pupils in non-selective state schools in England[1] ?
- To what extent does participation in SDI affect teacher workload?
- To what extent does participation in SDI affect Year 5 teachers' perceptions of their students' confidence in their Maths abilities.
- Does the impact of SDI on maths attainment of Year 5 pupils differ by FSM eligibility?
- What is the impact of SDI on the size of the gap between higher achieving and lower achieving Year 5 pupils?

Analysis will investigate the following primary hypothesis on an intention-to-treat basis.

Primary outcomes:

[1] Participating schools volunteered to take part in the study.

- H1: Participating in SDI improves Year 5 pupils' maths attainment, as measured in GL's Progress Test in Maths

Secondary outcomes:

- H2: Adopting SDI pedagogy and class structure reduces teacher workload, as measured by a survey of SDI teachers
- H3: Participating in SDI improves Year 5 teachers' perceptions of pupil's confidence regarding their maths abilities, as measured by a survey of SDI teachers

Sub-group effects:

- H4: SDI will have a different (higher or lower) impact on pupils ever eligible for Free School Meals (FSM) compared with those ineligible.

Additional analyses:

- H5: Pupils participating in the SDI will have a different (higher or lower) variance in attainment at follow-up.

Design Overview

Trial type and number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		Training hub (regional)
Primary outcome	variable	Year 5 student Maths attainment
	measure (instrument, scale)	GL's Progress Test in Maths
Secondary outcome(s)	variable(s)	-Teacher workload -Teachers perceptions of student self-confidence in maths
	measure(s) (instrument, scale)	-Teacher marking time, -Teacher perceptions regarding students' confidence in maths. These measures were collected at baseline and will be collected at endline using a bespoke teacher survey. For more details see Appendix 1.

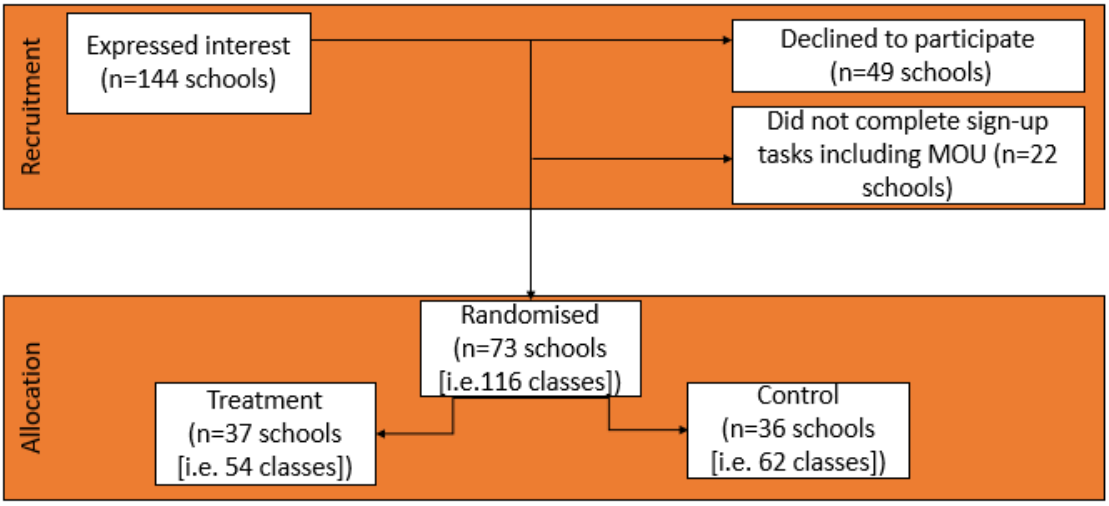
Follow-up

Below is a CONSORT flow-diagram outlining the flow of participating schools through the initial stages of the trial (Error! Reference source not found.). 144 schools were initially invited to take part in the trial, 49 of which declined to participate and a further 22 of which did not complete sign-up tasks such as completing the Memorandum of Understanding or enumerating pupils. This resulted in a total of 73 schools being included for randomisation.

After randomisation 37 schools were assigned to the treatment group and 36 schools to the control group.

This diagram will be updated in the final report to present a complete summary of the flow of trial schools and classes from recruitment through randomisation, post intervention assessment and analysis.

Figure 1: Consort diagram



Sample size calculations overview

		Protocol[2]		Randomisation[3]	
		OVERALL	FSM	OVERALL	FSM
MDES		0.27	0.3	0.28	0.32
Pre-test/ post-test correlations	level 1 (pupil)	0.5	0.5	0.5	0.5
	level 2 (class)	0.0	0.0	0.0	0.0
	level 3 (school)	0.1	0.1	0.1	0.1
Intracultural correlations (ICCs)	level 2 (class)	0.05	0.05	0.05	0.05
	level 3 (school)	0.14	0.14	0.14	0.14
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size (classes per school)		2	2	1.6 (1.4[1])	1.6 (1.4[1])
Average cluster size (pupils per class) [2]		27	4[3]	27	4[3]
	intervention	37	37	37	37

[2] As analysed
[3] At time of randomisation

Number of schools	control	36	36	36	36
	total	73	73	73	73
Number of pupils[4]	intervention	1,998	296	1,598[4]	237[4]
	control	1,944[2]	288[3]	1,555[4]	230[3, 4]
	total	3,942[2]	584[3]	3,153[4]	467[3, 4]

[1]Harmonic mean, based on data collected from participating Same Day Intervention schools
[2]We assume an average of 27 students per class. Figures based on data from Department for Education, Schools, Pupils and their Characteristics: January 2017 - National Tables
[3]Proportion of FSM students anticipated to be national average for age-group of 14.4%, as in Department for Education, Schools, Pupils and their Characteristics: January 2018 - National Tables. Totals rounded to nearest whole number.
[4]Number of pupils calculated using arithmetic mean number of classes per school (1.6), rather than harmonic mean

Randomisation was stratified by regional hub to allow for regional differences in implementation and school characteristics. For education programmes, the variance explained by pre-test scores can be relatively high if pre-test scores are used in adjusted analysis[4]. Our pre- and post-test measures are informed by DeMack, 2019[5], Torgerson and Torgerson (2013)[6] and Allen et al. (2018)[7]. School-level intra-cluster correlations (ICCs) are based on an EEF guidance note, using ICCs relating to Key Stage 2 Total Maths Scores for the North-West[8], while class-level ICCs are expected to be smaller.

Since writing the trial protocol, we estimated, using PowerUp![9] this study to be powered to detect an effect of 0.27 standard deviations based on the assumptions outlined in the first column of the Sample Size Calculations Table.

However, the mean number of classes per school for recruited schools was lower than anticipated (harmonic mean = 1.4 classes per school, rather than the anticipated 2). Column three of the Sample Size Calculations Table provides updated details regarding the calculation of our minimum detectable effects size. Assuming explanatory power of baseline scores of 50% at pupil and 10% at school level, and no further school level attrition or loss to follow-up of pupils, we estimate the study to be powered to detect an effect of 0.28 standard deviations[10].

Analysis

The evaluation of Same Day Intervention aims to evaluate its impact on the Maths attainment of Year 5 pupils in England and how it differs by FSM eligibility. The trial was designed as a two-armed, four-level randomised controlled trial. The highest level of

[4] Bloom, Howard S., Lashawn Richburg-Hayes, and Alison Rebeck Black. 2007. 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions'. Educational Evaluation and Policy Analysis 29 (1): 30–59.
[5] DeMack, S. 2019. Does the classroom level matter in the design of educational trials? A theoretical & empirical review. EEF Research Paper No. 003.
[6] Torgerson and Torgerson, 2013. Randomised trials in education: An introductory handbook. EEF
[7] Rebecca Allen, John Jerrim, Meenakshi Parameshwaran, Dave Thompson. Properties of commercial tests in the EEF Database. EEF Research Paper Series, No. 001, February 2018
[8] EEF, Intra-cluster correlation coefficients, 2015.
[9] Nianbo Dong and Rebecca Maynard, 'PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies', Journal of Research on Educational Effectiveness 6, no. 1 (1 January 2013): 24–67, doi:10.1080/19345747.2012.673143.
[10] These assumptions are in line with those set out in Torgerson and Torgerson (2013) and EEF (2013) for scenarios when the same pre-test and post-test are used.

clustering is the regional hub (Level 4) and Is controlled for with fixed effects. There are then three levels of random effects: Schools (Level 3), Classes (Level 2) and Pupils (Level 1).

As randomisation was stratified and our trial is nested, a single-level OLS approach to estimation and inference is not adequate as it does not account for correlation among observations within clusters, leading to underestimated standard errors. An alternative solution would be to use a single-level OLS and robust standard errors to account for non-independence between observations across clusters.[11]

A third option, and the one we suggest, is a multilevel model. The latter makes strong assumptions, (1) that random effects are normal, (2) that the model contains all relevant variables to assume that errors and regressors are uncorrelated at all levels, (3) that we have enough observations at each level. Single level OLS requires fewer assumptions. However, single-level analysis would not be able to identify what proportion of pupils' Maths attainment is attributable to school-level versus individual-level variation. Not accounting for clustering would produce downward biased standard errors and result in confidence intervals that are too narrow. Therefore, the primary analysis will use a multi-level model. A sensitivity analysis adopting a single-level OLS regression, using cluster robust standard errors, will also be estimated for the primary outcome. Further details on this sensitivity analysis is in the additional analysis section.

Primary outcome analysis

The main analysis will estimate the intervention's impact on enrolled Year 5 pupils' maths attainment, as measured by GL's Progress Test in Maths (raw scores), using an intention-to-treat approach. The test will be administered in May/June 2019 to all pupils that agreed to take part and have signed the MOU. Following EEF guidance, evidence of effectiveness and reported effect sizes will be obtained from a baseline-adjusted analysis, in which the dependent variable is the raw score of the GL Progress Test in Maths, and effects are estimated through a multilevel linear model containing a dummy variable indicator capturing treatment/control group membership, the stratification variable (i.e. regional hub), and pupil prior attainment (combination of Key Stage 1 [KS1] and Early years foundation stage profile [EYFSP]) in Maths at pupil level. Sensitivity tests will be conducted using an unadjusted analysis and an adjusted model with additional covariates, which are described fully in the section entitled *Additional Analyses*. This will be used to assess if there is a difference in impact estimates when controlling for potential baseline imbalance between the intervention and control groups.

There are clear limitations to using KS1 and EYFSP scores, not least because they are both categorical measures, limiting the amount of variance. An alternative would be to have conducted testing, however in addition to time and budget constraints, this could place an unnecessary burden on participating schools and pupils. To mitigate the low-variance, KS1[12] and EYFSP[13] mathematics scores will be combined (as a weighted sum) into a composite index, with greater weight placed on KS1 Maths (66%) than on EYFSP G11 (Numbers, 17%) and EYFSP G12 (Shape, spaces and measures, 17%). This weighting reflects the fact that KS1 scores are likely to be better predictors as they were conducted at a more recent time-point.

[11] Primo D., Jacobsmeier M., Milyo J., 2007, "Estimating the Impact of State Policies and Institutions with Mixed-Level Data", State Politics and Policy Quarterly, Vol. 7, No. 4: pp. 446-459
[12] The KS1_MATH_OUTCOME variable will be used.
[13] The FSP_MAT_G11 and FSP_MAT_G12 variables will be used.

The model analysed will account for four levels of clustering. The highest level of clustering is the strata of regional hubs. Nested within these regional hubs are three other levels: schools, classes and pupils. The highest level will be modelled as fixed effects as there are only three of them, the remaining clusters: school, class and pupils, will be random effects.

The basic form of the model is,

$$Maths\ Attainment_{ijk} = \beta_0 + \beta_1 Baseline_{ijk} + \beta_2 Intervention_k + \beta_3 Regional\ hub + u_{jk} + w_k + e_{ijk}$$

Where pupils (i) are clustered in classes (j) within schools (k). The intervention effect is estimated by β_2. β_3 represents the regional strata at randomisation and u_{jk} and w_k are respectively the class and school random-effect and e_{ijk} the error term. In line with the EEF Analysis Guidance, other covariates will not be considered at this stage. See later section for an explanation of how effect sizes will be calculated.

The analysis will be run in Stata 14 SE-64.

Optional primary outcome analysis

The Same Day Intervention evaluation also includes an option for follow-up analysis of pupils' maths attainment at Key Stage 2 (KS2). If this optional analysis were to be undertaken, this model would have the same specification, only changing the outcome of interest to KS2 maths scores from the NPD (MATMRK).

Secondary outcome analysis

The secondary outcome analysis will explore the impact the intervention has on teacher workload and on teachers' perceptions of pupil confidence. This secondary outcome analysis will use data from the teacher survey, a bespoke survey from NatCen to be administered at baseline and endline (see Appendix 1). For each survey question, responses will be combined as averages. Findings will be triangulated with those of the IPE where possible. The secondary outcome analysis will be estimated using a single-level model, rather than the multi-level model used for the primary analysis, as the number of teachers per school is likely to be too small to robustly estimate random effects[14].

The first of the two secondary analyses assesses teacher workload and has the following hypothesis:

- H2: Adopting SDI pedagogy and class structure reduces teacher workload, as measured by a survey of SDI teachers

Teacher workload is defined as: overall time spent marking Year 5 work from maths lessons, during and outside of lesson times, measured in minutes. Further details on the measurement of these outcomes are available in Appendix 1.

[14] In cluster samples, the design effect is approximately equal to 1+(average cluster size-1)*ICC. According to Muthén & Satorra, (1995), if the design effect is smaller than two, using single level analysis on multilevel data does not appear to lead to misleading results. In our case, with a maximum expected number of teachers per school being 3, the design effect would be 1+(3-1)*ICC. Assuming a conservative ICC of 0.2, the design effect would therefore be = 1+(2*0.20) = 1.4, a design effect smaller than 2.

The analysis of secondary outcomes will be conducted on an intention-to-treat basis using a single level OLS model using Huber-White cluster robust standard errors using the **robust** option of the **reg** command in STATA.

The basic form of the model is,[15]

$$Teachers'workload_{ij} = \beta_0 + \beta_1 Baseline_i + \beta_2 Intervention_i + \beta_3 Regional\ hub_j + e$$

Following EEF Analysis Guidance (EEF 2018), this includes baseline workload, a dummy variable identifying treatment allocation and the randomisation strata; regional hub.

In the equation above, (i) represents teacher level outcomes, and (j) the regional hubs level. The intervention effect is estimated by β_2, while β_3 represents the regional strata at randomisation and e the error term[16]. This model assumes that the majority of teachers surveyed at endline are the same as those surveyed at baseline. It may be that between baseline and endline, some teachers may have changed roles within schools, or left schools entirely. If a majority of teachers in the endline survey were not also surveyed at baseline, this will mean a lower correlation between baseline and endline, thereby reducing the explanatory power of the baseline score as a covariate. We will run a sensitivity analysis that excludes $\beta_1 Baseline_i$ from the model to compensate for possible teacher turnover and response rate. The results of both analyses will be interpreted cautiously in terms of their generalisability.

The second of the secondary analyses assesses teacher's perception of pupils' confidence in maths has the following hypothesis:

- H3: Participating in SDI improves Year 5 teachers' perceptions of pupil's confidence regarding their maths abilities, as measured by a survey of teachers

Teacher perceptions of pupil confidence are measured using questions 2a and 2b in the teacher survey, outlined in Appendix 1, comparing the current cohort of pupils with that for a previous cohort of pupils. Descriptive analysis of these responses at baseline and at follow-up will compare the proportions in the treated and control groups, tested for significance with a Chi-square test.

Analysis for H3 will follow a similar method to that for H2, a single-level OLS model using cluster robust Huber-White standard errors. The analysis model will take the following form:

$$Pupil\ confidence_{ij} = \beta_0 + \beta_1 Baseline_i + \beta_2 Intervention_i + \beta_3 Regional\ hub_j + e$$

Interim analyses

No interim analyses are planned for this trial.

Subgroup analyses

The subgroup analyses will explore the following hypotheses:

[15] This assumes that the majority of teachers surveyed at endline are the same as those surveyed at baseline. It may be that between baseline and endline, some teachers may have changed roles within schools, or left schools entirely. If a majority of teachers in the endline survey, were not also surveyed at baseline, including $\beta_1 Baseline_i$ may not improve explanatory power.

[16] Huber-White robust standard errors will be calculated using the 'robust' command in Stata.

- H4: SDI will have a different (higher or lower) impact on pupils eligible for Free School Meals (FSM) compared with those ineligible (assessed using *ever FSM* (EVERFSM_6_P) from NPD)

Subgroup impacts on the primary outcome will be estimated for pupils eligible for FSM (EVERFSM_6_P). This will involve the re-estimation of the model described in the primary outcome section with the addition the FSM indicator and an interaction term combining FSM eligibility and treatment allocation. Where the coefficients resulting from this interaction reach statistical significance at the 95% level, separate models will be estimated and reported for each subgroup. The trial will likely not provide sufficient power to fully explore this, so these results are likely to be only indicative.

Additional analyses

We will undertake the following exploratory additional analysis to explore whether the Same Day Intervention has had a significant impact on the attainment gap within classes.

- H5: Pupils participating in the SDI will have a different (higher or lower) variance in attainment at follow-up

The distributions of the outcome will be displayed graphically and two statistical tests to measure the dispersion of scores between the treated and control group will be undertaken. These tests are Levene's test (Levene, 1960) and the Brown-Forsythe test (Brown & Forsythe 1974). These two tests can be used to test the equality of standard deviation of two groups. Levene's test explores the equality in standard deviation between groups at the mean, whilst the Brown-Forsythe test examines this at the median[17].

The test statistics and their associated P-values will be presented as indicative analysis only and will be triangulated with findings from the process evaluation. These statistics will be estimated using the **robvar** command in STATA, a robust test for the equality of variances. This hypothesis will be further explored through the process evaluation of the Same Day Intervention.

A range of sensitivity analyses will also be carried out as additional analyses to explore the robustness of the main findings, with findings for all models transparently reported. If a sensitivity analysis finds any substantively different finding to the main analysis, this will be acknowledged. The following analyses will be carried out:

- An unadjusted analysis that will not include baseline covariates;
- An adjusted model, including a wider range of prognostic covariates to control for potential imbalance at baseline: free school meal eligibility, gender and school type
- A single-level OLS regression model, using Huber-White cluster-robust standard errors. The variables included will be the same as the primary analysis model: baseline attainment, treatment allocation and regional hub.

Imbalance at baseline

Randomisation, if conducted correctly, should result in there being no important differences between treatment and control groups in the main determinants of our outcomes of interest. Any such differences arising will do so by chance. We will explore the potential for chance

[17] This tends to be more robust if the distribution is skewed.

imbalances first through an inspection of the descriptive statistics of various characteristics, comparing treatment and control groups. Baseline characteristics will be summarised by treatment and control group across schools and pupils. Where available variables will be presented at pupil level, otherwise at school level.

Continuous variables will be summarised with descriptive statistics (n, mean, standard deviation, range, median and as effect sizes).

At school level, the comparison will cover:

- School type, (NFTYPE)

At pupil level, the following baseline comparisons will be presented:

- Ever received FSM
- Gender
- Key Stage 1 and EYFSP combined scores

Imbalance on baseline covariates between the treatment and control groups in the sample as analysed will be assessed for the covariates listed above using the appropriate statistical test (two-independent-sample *t*-test for continuous variables and Fisher's exact test for categorical variables). Hedge's g effect sizes will also be estimated, with an effect size of greater than 0.05 considered as an indication of possible imbalance.

If imbalances are indicated, a model that includes the unbalanced variables (i.e. where Hedge's g is greater than 0.05) in addition to those in the main model will be estimated as a sensitivity analysis.

Missing data

For the main primary analysis of pupil outcomes, it is possible that there may be loss to follow up due to moves and other external factors influencing participation in the final outcome testing[18]. Baseline data will be sourced from the National Pupil Database (NPD). Very-low levels of attrition may occur if the pupils cannot be linked, or if they are missing baseline data. Given the available information on the retention of schools to the trial, we anticipate very small attrition at school level.

For the secondary analysis, we may experience a larger loss due to low and potentially differential response rates between the treatment and control group to the follow-up teachers survey.

For the primary and secondary analyses, we will assume that missing data are missing completely at random and use complete case analysis. We will then conduct sensitivity analyses to assess the robustness of the inferences about treatment effects to alternative assumptions about the mechanisms leading to missing data.

We will explore the extent and pattern of missingness for both primary and secondary outcomes if missing data exceeds 5 percent[19]. First, to explore the extent of missingness, the number of pupils/ teachers with missing outcomes will be reported by treatment status. Additionally, baseline comparisons between pupils/teacher of each treatment arm will be

[18] However, as we use unique identifiers (e.g. names and date of birth), to link NPD data to our pupils, a small loss can occur due to erroneous data given by students such as date of birth.
[19] In line with EEF Analysis Guidance, 2018

compared with observed and missing values using cross-tabulations. Secondly, to explore the pattern of missingness, we will estimate a logistic regression with loss to follow-up as a binary outcome, and covariates as potential predictors of missingness[20]. These covariates will include all the characteristics explored for baseline balance (free school meal eligibility, gender and school type) baseline attainment) and other characteristics available in the NPD data.

If any covariate can predict loss to follow-up, we will conduct sensitivity analyses under the assumption that outcome data are missing at random. We will use multiple imputation to infer the likely results of those lost to follow-up and present results alongside headline impact estimates for comparison. The model will include all variables in the adjusted analysis: treatment allocation, baseline attainment, treatment allocation, randomisation strata, free school meal eligibility, gender and school type. This will generate predicted values for the missing cases and estimate treatment effects and standard errors under this alternative assumption. However, if loss to follow-up cannot be predicted using existing covariates, multiple imputation will not be possible). The implication of this will be discussed clearly in the final report.

Compliance

Whilst Intention-to-Treat (ITT) analysis is informative to policymakers about the effects of an *offer* of treatment, it is not informative about the impact of an intervention on those who receive it. Consequently, the trial analysts propose conducting analysis of non-compliance. There are several potential areas for non-compliance issues in this trial. Non-compliance could arise because of:

- Staff not attending training sessions
- Teachers not delivering the programme as intended
- Schools assigned to control delivering the intervention

Compliance will be measured at the school level for both treatment and control schools.

The evaluation will collect compliance data on two types of measure of compliance:

Attendance at training sessions was identified in the Theory of Change (TOC) as a key part of the intervention. Teachers, Headteachers are required to attend some training. Teachers are required to attend three full-day training sessions, whilst Headteachers are required to attend just one full day's training. Training is optional for Teaching Assistants and there are also additional 'twilight' sessions that teachers can attend if they wish, but these are not a compulsory part of the intervention and consequently will not be considered in a measure of compliance.

For treatment schools, attendance at training has been recorded using registers of training sessions collected by the delivery partner.

Fidelity: In addition to attending training, it is also important to capture whether the training was delivered as intended (fidelity). The *Same Day Intervention* has five key elements:

[20] We will test for multicollinearity and address where necessary by removing collinear variables.

1. Use of Same Day Intervention pedagogical techniques to model new concepts at the start of each Same Day lesson
2. Re-structuring the maths lesson to an hour and fifteen minutes, including a 15 minute 'pitstop'
3. Teachers marking an assessment whilst pupils are out of the classroom
4. Availability of a Teaching Assistant for all Same Day Intervention classes
5. Splitting the class into two-groups based on the results of the assessment, with the teacher teaching the group in need of more support whilst the teaching assistant teaches the other group

The endline teacher survey will collect data on whether each of these elements was a part of maths classes in treatment schools, with teachers asked to report whether, over the past year, each element was incorporated,

- Always
- Regularly
- Occasionally
- Not at all

Approach to compliance

We will run descriptive statistics on both types of measure. The compliance analysis will use an index combining both measures of compliance as this represents the actual nature of the program delivered by schools. Compliance will be measured at the school level, as follows,

We will construct an index in which attendance at training and intervention fidelity are given equal weighting, with attendance captured via training registers and fidelity captured via the post-intervention survey of teachers. Attendance will be summarised for each class in each school as a proportion of all compulsory training sessions attended by teachers, where total possible sessions is equal to three (i.e. the three sessions each class teacher was invited to attend). An average will then be taken across all classes within a school. Whether or not the Headteacher attended training will then be added to the teacher average, allowing a total score between zero and four.

For fidelity, the categorical variables listed above will be combined into a single continuous variable capturing each of these elements, where a value of zero indicates "Not at all" and a value of three indicates "Always". The scale will then have a possible range of zero to 15. The attendance and fidelity scales will then be combined into an index with a possible range of $0 \leq Comply^{T} \leq 15$

The two scores (attendance and fidelity) will then converted to provide an overall measure of compliance with a range from zero to one, with each score given equal weight.

Thus-far, the discussion has focused on non-compliance in the treatment arm. At the time of writing, the trial analysts are aware of two-sided non-compliance. That is, at least one school assigned to the control-arm have accessed a version of the *Same Day Intervention*. Compliance in control schools will be measured using the same index as outlined above.

The post-intervention survey will collect information on whether this intervention was delivered in control schools in the same way as it does for treatment schools. Since none of the control schools attended any version of the Same Day Intervention training, the school survey will not ask about attendance at Same Day training and control schools will be automatically awarded a score of 0 for the training component of the compliance measure. Control schools will be asked in the school survey whether they have implemented any of the Same Day intervention components outlined in the Fidelity measure above and will be awarded a score of up to 15 for fidelity to the Same Day Intervention. The training and fidelity scores will be combined into an index following the same principles as outlined above.

Due to the two-sided non-compliance, the compliance analysis will estimate a Local Average Treatment Effect (LATE)[21]. This LATE will be estimated using an instrumental-variable (IV) approach (2SLS) using random assignment as the instrument, in line with Angrist and Imbens (1995). The first stage equation is as follows:

$$Comply_j = \propto + \beta_1 Treat_j + \varepsilon_{ij}$$

The predicted values of compliance $\widehat{Comply_j}$ will then be used in the estimation of the second stage model, as follows:

$$Y_{ij} = \propto + \beta_1 Treat_j + \beta_2 \widehat{Comply_j} + \beta_3 Baseline_{ij} + Hub_{ijk} + \omega_{ij}$$

Where $Baseline_{ij}$ indicates prior attainment and Hub_{ijk} indicates the stratification at randomisation on regional hubs, with ω_{ij} representing the error term. The coefficients β_2 and β_3, will be used to calculate the LATE. To ensure correct estimation of standard errors with clustered data, the model will be estimated with cluster robust standard errors and using **ivregress** in STATA. In line with EEF guidance (EEF, 2018) the correlation between the instrument $Treat_j$ and the endogenous variable will be reported along with the F-statistic.

Intra-cluster correlations (ICCs)

The intra-cluster correlations (ICCs) will be calculated directly from the primary analysis model, using the variance estimates for each level of clustering. The formula used to calculate the ICC for schools $\boldsymbol{\rho_S}$, and classes $\boldsymbol{\rho_C}$, is as follows:

$$\boldsymbol{\rho_S} = \frac{\boldsymbol{\sigma^2_{BS}}}{\boldsymbol{\sigma^2_{BS}} + \boldsymbol{\sigma^2_{BC}} + \boldsymbol{\sigma^2_{WC}}} = \frac{\boldsymbol{\sigma^2_{BS}}}{\boldsymbol{\sigma^2_{WT}}} \qquad (1)$$

$$\boldsymbol{\rho_C} = \frac{\boldsymbol{\sigma^2_{BC}}}{\boldsymbol{\sigma^2_{BS}} + \boldsymbol{\sigma^2_{BC}} + \boldsymbol{\sigma^2_{WC}}} = \frac{\boldsymbol{\sigma^2_{BC}}}{\boldsymbol{\sigma^2_{WT}}} \qquad (2)$$

In these formulae $\boldsymbol{\sigma^2_{BS}}$ represents the between-school variance, $\boldsymbol{\sigma^2_{BC}}$ the between-class variance, $\boldsymbol{\sigma^2_{WC}}$ the within-class variance and $\boldsymbol{\sigma^2_{WT}}$ the sum of the variance at all levels. The ICCs will be calculated using the STATA package **estat icc.**

[21] The interpretation of the LATE is different to that of Complier Average Causal Effects (CACE). As a result, this effect size must be interpreted as the average effect on compliers.

Effect size calculation

The impact estimates will be reported as Hedges' *g* effect sizes. Hedges (2011) constructed formulae for effect sizes for three-level[22] cluster randomised trials, though these do not account for covariate adjustment. These formulae have therefore been adjusted by the trial analysts, following the approach of Borenstein (2009). This results in the difference in adjusted means being scaled by the pooled sample variance of the post-test measures (i.e. the unadjusted variance). The analysts will use the formulae outlined below, using 95% confidence intervals.

The point estimate, *g*, is calculated as the difference between adjusted group means $\bar{Y}^T_{adj}$ and $\bar{Y}^C_{adj}$, scaled by the unconditional total standard deviation within-treatment groups S_{WT}, and adjusted to account for school and class-level clustering, as follows:

$$g_{WT} = J \times \left(\frac{\bar{Y}^T_{adj} - \bar{Y}^C_{adj}}{S_{WT}}\right)\sqrt{1 - \frac{2(p_U - 1)\rho_S + 2(n_U - 1)\rho_C}{N-2}} \qquad (3)$$

Where *J* is the bias correction to estimate Hedges' *g* from Cohen's *d*, given by:

$$J = 1 - \left(\frac{3}{4(n_T + n_C - 2) - 1}\right) \qquad (4)$$

The standard deviation is the square root of the estimated pooled variance, S^2_{WT}, calculated as:

$$S^2_{WT} = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\sum_{k=1}^{n_{ij}^T}\left(Y^T_{ijk} - \bar{Y}^T\right)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\sum_{k=1}^{n_{ij}^C}\left(Y^C_{ijk} - \bar{Y}^C\right)^2}{N-2} \qquad (5)$$

In these formulae, the subscripts i, j and k represent pupils, classes and schools respectively.

The school intra-cluster correlation, ρ_S and the class intra-cluster-correlation, ρ_C are given by the formulae (1) and (2) in the previous section. The remaining terms are calculated as follows:

$$p_U = \frac{N^C \sum_{i=1}^{m^T}\left(\sum_{j=1}^{p_i^T} n_{ij}^T\right)^2}{NN^T} + \frac{N^T \sum_{i=1}^{m^C}\left(\sum_{j=1}^{p_i^C} n_{ij}^C\right)^2}{NN^C} \qquad (6)$$

$$n_U = \frac{N^C \sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\left(n_{ij}^T\right)^2}{NN^T} + \frac{N^T \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\left(n_{ij}^C\right)^2}{NN^C} \qquad (7)$$

$$N = N^T + N^C = \sum_{i=1}^{m^T}\sum_{j=1}^{P_i^T} n_{ij}^T + \sum_{i=1}^{m^C}\sum_{j=1}^{P_i^C} n_{ij}^C \qquad (8)$$

The 95% confidence intervals will be calculated as follows:

$$g_{WT} - 1.96v_g \leq \delta_T \leq g_{WT} + 1.96v_g \qquad (9)$$

The variance of the effect size estimate, v_g, can be conservatively approximated by:

[22] Although this analysis accounts for four levels of clustering, the highest level is controlled for using fixed effects, the three-levels of random effects therefore make these formulae appropriate. The equations below are adapted from equation 31 in Hedges (2011).

$$v_{\{g_{WT}\}} = \frac{(1+(p_U-1)\rho_s+(n_U-1)\rho_C)(1-r^2)}{\tilde{N}} + \frac{d_{WT}^2}{2(M^T+M^C-2)-q`-1} \quad (10)$$

Where r^2 is the covariate outcome correlation, $q`$ the number of covariates, M^T and M^C the number of schools in the treatment and control groups respectively. Finally, $\tilde{N}$ is given by:

$$\tilde{N} = \frac{N^T N^C}{N^T+N^C} \quad (11)$$

References

Angrist, J., & Imbens, G. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. American Statistical Association, 90(430), 431-442.

Michael Borenstein, “Effect Sizes for Continuous Data”, in *The Handbook of Research Synthesis and Meta-Analysis*, ed. Harris M. Cooper, Larry V. Hedges, and Jeffrey C. Valentine, 2nd edition, (New York, Russell Sage Foundation), 2009, pp. 221-236

Brown, M. and Forsythe, A. (1974) Robust Tests for the Equality of Variances, *Journal of the American Statistical Association*, 69(346), 364-367.

Education Endowment Foundation (2013) Pre-testing in EEF evaluations.

Education Endowment Foundation (2018) Statistical analysis guidance for EEF evaluations, https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf. [Accessed 06/12/18]

Larry V. Hedges, “Effect Sizes in Three-Level Cluster-Randomized Experiments” *Journal of Educational and Behavioural Statistics*, 36 (3), 2011, pp.346-380, doi: 10.3102/1076998610376617, equation 31 and following.

Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. In P.V.Marsden (Ed.), Sociological methodology (pp. 267-316). Oxford, England: Blackwell.

Nianbo Dong and Rebecca Maynard, ‘PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies’, Journal of Research on Educational Effectiveness 6, no. 1 (1 January 2013): 24–67, doi:10.1080/19345747.2012.673143.

Levene, H. (1960). Robust tests for equality of variances, In “Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling” (Olkin, I., Ghurye, S., Hoeffding, W., Madow, W. & Mann, H. eds.). Stanford University Press, 278–292.

Torgerson, C. & Torgerson, D. (2013). Randomised Controlled Trials in Education: An Introductory Handbook. Educational Endowment Foundation.

Appendix 1: Teacher Workload Survey

This is a short survey for **all current Year 5 teachers in your school**. Completing this will help us to gather information the time you spend marking Year 5 work from maths lessons.

This survey must be completed by 27th April 2018. Failure to do so will mean your school will not be included in the trial.

If you have any questions when completing this survey, please contact the NatCen team directly on 0808 169 5668 or email sameday@natcen.ac.uk

Survey Questions: outcomes

Please complete all questions.

Q1a. Teacher workload:a {Ask all}

This question is about how much time you spend overall on marking Year 5 work from maths lessons, including time spent marking work inside of lesson time.

Baseline

During this academic year (2017-18), in an average week, how much time (in minutes) do you spend marking Year 5 work from maths lessons?

Endline

During this academic year (2018-19), in an average week, how much time (in minutes) did you spend marking Year 5 work from maths lessons?

Q1b. Teacher workload:b {Ask all}

This question is about how much time you spend overall on marking Year 5 work from maths lessons, excluding time spent marking work inside of lesson time.

Baseline

During this academic year (2017-18), in an average week, how much time (in minutes) do you spend marking Year 5 work from maths lessons, outside of lesson time?

Endline

During this academic year (2018-19), in an average week, how much time (in minutes) did you spend marking Year 5 work from maths lessons, outside of lesson time?

Q2. Student confidence {Ask all}

Asked at Endline only

The next two questions are about your perception of students' confidence in their maths abilities. Please compare this year's Year 5 cohort (2018-2019) to last year's Year 5 cohort (2017-2018).

Q2a. As far as you aware, how does this year's Year 5 cohort that you teach compare to last year's Year 5 cohort?

In regards to their confidence in maths, compared to last year's cohort, this year's cohort are;

A lot more confident, a little more confident, about the same, a little less confident, a lot less confident

Q2b. Please think of the lowest achieving students in the Year 5 cohort you currently teach. How do they compare to the lowest achieving students from last year's Year 5 cohort?

In regards to their confidence in maths, compared to last year's cohort, this year's cohort are;

A lot more confident, a little more confident, about the same, a little less confident, a lot less confident

---