Statistical Analysis Plan
Maths-for-Life

Evaluator (institution): Behavioural Insights Team
Principal investigator(s): David Nolan

PROJECT TITLE	Maths-for Life
DEVELOPER (INSTITUTION)	University of Nottingham
EVALUATOR (INSTITUTION)	Behavioural Insights Team
PRINCIPAL INVESTIGATOR(S)	David Nolan
TRIAL (CHIEF) STATISTICIAN	David Nolan
SAP AUTHOR(S)	David Nolan, Patrick Taylor
TRIAL REGISTRATION NUMBER	ISRCTN14810016
EVALUATION PROTOCOL URL OR HYPERLINK	https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-for-life/

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1	10 June 2020	COVID contingency updates, removal of Cohort 2 analysis
1.0	04 March 2019	N/A

Table of contents

Introduction

The Maths-for-Life intervention aims to improve maths GCSE outcomes for students resitting their exams. It does this by creating a more student-centred classroom, focussed on problem solving and dialogic teaching. The intervention will be evaluated by a two-arm, cluster randomised controlled trial, with randomisation at the "setting" level. We define settings as any institution in which students may resit their Maths GCSE exam; settings therefore include Further Education (FE) Colleges, 6th Form Colleges, Schools, and Training Providers. The primary outcome of the evaluation is KS5 GCSE Maths resit performance, as indicated by a student's Uniform Mark Scale (UMS) score. The effect of the intervention on two secondary outcomes will also be estimated: GCSE maths pass rate and mathematical self-efficacy. Further analysis will also be carried out to estimate the Complier Average Causal Effect. All effects will be estimated for the whole sample and for the subgroup of pupils who were eligible for free school meals (FSM) in the year of the trial and/or the previous six years ('Ever 6 FSM') to assess whether the intervention has differential effects for disadvantaged students.

Design overview

Trial type and number of arms		Two-arm, cluster randomised
Unit of randomisation		Setting (FE college/6th Form College/School/Training Provider, or a sub-site of one of these)
Stratification variables (if applicable)		Setting type
Primary outcome	variable	Maths attainment
	measure (instrument, scale)	GCSE Maths UMS score[1]
Secondary outcome(s)	variable(s)	KS5 Maths GCSE pass rates, Mathematical self-efficacy
	measure(s) (instrument, scale)	KS5 Maths GCSE grade, Part E of the Year 10 Teleprism survey

Follow-up

NA

[1]Discussions with both exam boards and settings has provided confirmation that students' raw marks are supplied to each setting. The Joint Council for Qualifications (JCQ) steering group has provided a framework to convert raw marks onto a Uniform Mark Scheme (UMS), adjusting for exam board and tier of paper, which BIT researchers will implement.

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.15	0.22	0.23	0.29
Pre-test/ post-test correlations	level 1 (pupil)	0.5	0.5	0.5	0.5
	level 2 (class)[2]	NA	NA	NA	NA
	level 3 (school)[3]	NA	NA	NA	NA
Intracluster correlations (ICCs)	level 2 (class)[4]	NA	NA	NA	NA
	level 3 (school)	0.2	0.2	0.2	0.2
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		80 students per college, 10 per school	24 students per college, 2 per school	73 students per college, 32 per school/ 6th form college/ Training provider	9 students per college, 5 students per school/ 6th form college/ Training provider
Number of FE colleges	intervention	25	25	33	33
	control	25	25	32	32
	total	50	50	65	65
Number of schools / 6th Form College / School / Training Provider	intervention	30	30	17	17
	control	30	30	18	18
	total	60	60	35	35
Total number of settings	intervention	55	55	50	50
	control	55	55	50	50

[2] This is not estimated as the majority of students participating in the study have moved to a new educational institution and we are not administering any pre-test assessment

[3] This is not estimated for the same reasons as above

[4] As our analytical strategy clusters our standard errors at the level of randomisation (Setting), we only estimate ICCs at this level. We do appreciate that for settings with multiple classes there may be additional clustering at this level. However, we use a conservative estimate of setting level ICC to adjust for any clustering at the class level and include a Hierarchical Linear Model (HLM) which includes class level and setting level error terms as a robustness check for our analysis. Additionally, anecdotal evidence suggests students switch classes throughout the year, and class composition can change which would dilute any clustering effect at this level.

	total	110	110	100	100
Number of pupils	intervention	2300	660	3240	409
	control	2300	660	2735[5]	340
	total	4600	1320	5975[6]	749

Prior to recruitment (at the stage of writing the Trial Protocol) we used a statistical process known as simulation-based inference to conduct all power analyses, using the R statistical software package. This was because we intended to recruit different types of educational settings with substantial variation in cluster size and large variation in cluster size tends to reduce statistical power in cluster randomised controlled trials[7]. Often, researchers will try to factor this into their power analysis by an inflation factor known as the coefficient of variation, incorporating this into the closed-form mathematical equations which are commonly used to derive statistical power for simple trials.

In our case, the combination of setting randomisation and the presence of many 'types' of settings with quite different properties necessitated using a simulations based approach as our MDES estimates incorporated the issues presented by varying cluster sizes[8]. Our simulations used a Poisson distribution to estimate cluster sizes based on the assumptions made regarding mean cluster size in each setting type (specified in Protocol section of the table above). However, post-recruitment cluster sizes are known, therefore we have deferred back to using closed-form algebraic equations to conduct our power analysis, and validate the results from the simulation-based approach taken in the trial design phase.

Please note that our MDES estimate has increased slightly compared to when estimated at the stage of the trial protocol. It must be noted we recruited fewer clusters than specified in the pre-trial power analyses, and are now using a different approach to conduct power analysis which represents our best current estimate of statistical power. It must be noted that we collected FSM status directly from the settings, therefore the estimate provided above is based on this data. However, we will estimate the MDES for FSM students again after access to the National Pupil Database (NPD) is granted and we can merge our sample data with NPD variables, as during the collection process settings flagged that FSM status was difficult to provide and they were unsure of the accuracy of their data.

Justification for the assumptions in the table above are as follows:

- **Pre-test/post-test correlations:** KS4 mathematics grade & KS2 mathematics score, have correlation 0.5 with our KS5 mathematics GCSE outcome.[9]

[5] The difference in total students between treatment and control groups is driven by a higher average number of students per setting in the treatment group (65) relative to the control group (55). However, we have conducted balance checks and this difference is not statistically significant.
[6] This figure may change as some students may pass in the November resit, results are released in January 2019. Any students who do pass will be excluded from the trial. This has been factored into our power analysis using a student attrition rate of 20%.
7 Lauer, S. A., Kleinman, K. P., & Reich, N. G. (2015). The effect of cluster size variability on statistical power in cluster-randomized trials. PloS one, 10(4), e0119074.
[8] The user-written code was validated by comparing results to those returned by closed-form mathematical equations used for simpler designs. It was quality assured by BIT's Chief Scientist and a commented version will be provided in the final report.
[9] Given the literature in this domain is quite sparse, there were no formal correlations available to use. We have made a conservative estimate revising down the KS2 and KS4 mathematics examination

- **Alpha and Power:** These are standard assumptions.
- **One or two-sided test:** A two-sided test is performed to err on the side of caution. There is limited existing evidence relating to the effect of interventions of this type on academic attainment, and the pilot evaluation – whilst broadly positive – revealed some risks associated with scaling, so we cannot assume the direction of any effect that we might observe.
- **Number of FE colleges:** 65
- **Number of schools/6th form colleges/training providers:** 35
- **Number of students per FE college:** Known (73 on average)
- **Number of students per school/6th form college/training provider: Known (**32 on average)
- **Student attrition**: 20% (in line with EEF guidelines)
- **Baseline resit pass-rate**: 28%[10]
- **Free School Meals:** Known **(**9 on average per college**,** per school/ 6th form college/ Training provider)
- **Simulated treatment effect:** When conducting power analysis it was assumed to be uniform across all participants in the treatment group. Although this assumption is simplistic and may be unrealistic, there is not, ex ante, any empirical reason to assume any particular different functional form for the treatment effect.

Analysis

The analysis plan is described in the sections that follow. All analyses will be carried out using the statistical software Stata (version 14). See Appendix 1 for the associated (indicative) Stata code that will be used.

Primary outcome analysis

Outcome

The primary outcome is overall KS5 mathematics GCSE resit performance for the academic year 2018/2019, as measured by the UMS score.

Analysis

Primary analysis will be intention-to-treat (ITT), in which we test the hypothesis that being assigned a place on the programme has an effect on performance. Analysis will be carried out using an OLS regression, specified below. Please note a class subscript, c, has been added to highlight that student are grouped within classes within settings, however treatment allocation is determined at the setting level.

$$Y_{ics} = \beta_0 + \beta_1.T_i + \alpha \mathbf{X}_{is} + \varepsilon_{ics}$$

where:

- Y_{ics} is the outcome for the KS5 mathematics GCSE resit performance for individual i, in class c, in setting s, measured by UMS score;

correlations from 0.7 (the correlation between KS2 & KS4 mathematics examinations) to 0.5 to account for this.

[10] Thomson, D. (2017) GCSE results day 2017: Good news about resits in English. Education Datalab. Available online: https://educationdatalab.org.uk/2017/08/gcse-results-day-2017-good-news-about-resits/ [Last accessed: 30 April 2018]

- T_i is a binary indicator for the treatment assignment for individual i (1 if the student is assigned to treatment and 0 if not);
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (FE College, 6th Form College, School, Training Provider[11]), baseline attainment (measured through both KS2 raw mathematics scores and KS4 mathematics GCSE grade), and a dummy variable for whether a student was a recipient of Free School Meals; and
- ε_{ics} is the cluster-robust error term, for individual i, in class c, in setting s, clustered at the setting level (assuming the errors are correlated within setting and reflecting the design of the study).

While UMS scores are bounded between 0 and 100, we assume that the response to the treatment will be locally linear so an OLS will be appropriate (in any case OLS gives the best linear approximation). Additionally, our estimated baseline is 28%.

We do note that the above model fails to account for any class-level clustering effects. We have elected to use an OLS regression as opposed to random effects specifications as random effects models require strict exogeneity of the regressors with the error term. Additionally, anecdotal evidence suggests students switch classes throughout the year, and class composition can change which would dilute any clustering effect at this level. However, we do note that not accounting for class-level clustering may impact our standard errors, therefore we have chosen to use a Hierarchical Linear Model (HLM) as a robustness check, An HLM incorporates the multilevel clustered nature of the data into the specification, and we will use results from our HLM to verify our primary analysis model choice.

Secondary outcome analysis

Outcomes

Effects will be estimated for two secondary outcomes for the academic year 2018/2019:

- **KS5 mathematics GCSE pass rates,** as measured by a binary outcome variable indicating whether or not a student achieved a grade 4 or higher. This will be obtained from the National Pupil Database (NPD) using the KS4_L2BASICS_94[12] variable.
- **Mathematical self-efficacy,** as measured by Part E of the Year 10 Teleprism survey[13]. This survey asks students to rate how confident they are - on a four-point Likert scale that ranges from 'Not confident at all' to 'Very confident' - in answering a range of questions that correspond to the following GCSE Maths topics: Number, Algebra, Geometry and measures, Ratio, proportion and rates of change, and Statistics. The total score from the survey for each student will be used in the analysis.

For the analysis of these outcomes, we will replace Y_i in our specification of the primary analysis with variously:

- GCSE mathematics pass or fail binary measure (1 if grade 4 or above, 0 if not)
- Mathematical self-efficacy score

[11] The category 'Training Provider' was not included in the trial protocol as it was not anticipated, but the developer has recruited a small number of these settings so we have added it here.
[12] The KS5 resit examination overwrites the KS4 Maths grade in the NPD tables
[13] For information on the design and validation of this surveying approach see: Pampaka, M., Kleanthous, I., Hutcheson, G. D., & Wake, G. (2011). Measuring mathematics self-efficacy as a learning outcome. Research in Mathematics Education, 13(2), 169-190. The survey itself is available online at: http://www.teleprism.com/surveys.htm. [Last accessed: 30 April 2018].

We have elected to fit a linear model as opposed to a logistic regression for our binary outcome measure (GCSE mathematics pass or fail) as our baseline pass-rate is 28%[14]. At this baseline, a linear model will approximate the results returned from a logistic regression and has the advantage of easier interpretation of parameter estimates.

Interim analyses

NA

Subgroup analyses

We will conduct analysis on the primary outcome for the subgroup of pupils who are registered for free school meals in the NPD (using the EVERFSM_6_P variable), using the same model as our primary analysis, with the addition of an interaction between treatment assignment and FSM status, to assess whether there is a significant difference in the treatment effect between FSM students and others.

We will also estimate the treatment effect for the subsample of participants who have been eligible for FSM in the past 6 years (EVERFSM_6_P = 1) and compare this to the estimated treatment effect for those not eligible for FSM (EVERFSM_6_P = 0). This will be operationalised using the regression model in the primary analysis for each of these two groups. For any mature learners (aged 19+) in our sample, the EVERFSM_6_P measure is not defined (i.e. is always set to 0). Therefore, we will exclude mature learners from this subgroup analysis.

We will also conduct analysis on the primary outcome by setting type, using the same model as our primary analysis, with the addition of an interaction between treatment assignment and setting type, to assess whether there is a significant difference in the treatment effect between students in different setting types.

Additional analyses

Robustness Checks

We will conduct robustness checks for all primary and secondary analysis using a Hierarchical Linear Model (HLM) which is an augmented OLS specification that takes into account the hierarchical structure of the data. Given that our outcome data is “nested” in that students are clustered in an organised manner (classes within settings), we are conducting this analysis to check whether including this in our estimation strategy effects our results. We can compare estimates produced by our primary analysis strategy and our HLM to examine the extent to which this is the case. As the HLM requires strict exogeneity of the random effects and our fixed predictors, we will follow guidance from Rice, Jones and Goldstein (2002)[15] to ensure our estimation strategy is asymptotically consistent.

[14] Thomson, D. (2017) GCSE results day 2017: Good news about resits in English. Education Datalab. Available online: https://educationdatalab.org.uk/2017/08/gcse-results-day-2017-good-news-about-resits/ [Last accessed: 30 April 2018]

[15] Rice, N., Jones, A. M., & Goldstein, H. (2002). Multilevel models where the random effects are correlated with the fixed predictors. *Centre for Health Economics, University of York, York*.

The multilevel regression model specification is below (we cannot include school or class fixed effects due to randomisation at the setting level). We plan to estimate a random intercepts (i.e. average student scores vary by class and setting) rather than a random slopes (i.e. treatment effect varies by class and setting). We will however check using a Bayesian Information Criterion (BIC) approach whether a random slopes model improves the model fit and if so we will switch to a random slopes model. However, as random slopes would decrease our degrees of freedom we have opted at this stage for a random intercepts model.

We do note that anecdotal evidence suggest students switch classes throughout the year, and class composition can change which limits the accuracy in estimation the class-level error term.

$$Y_{ics} = \beta + \beta_1 T_i + \alpha \boldsymbol{X}_{is} + \varepsilon_{ics}$$

$$\beta = \beta_0 + \gamma_{cs}$$

$$\gamma_{cs} \sim N(0, \sigma_c^2 + \sigma_s^2)$$

where:

- Y_{ics} is the outcome for the KS5 mathematics GCSE resit performance for individual i, in class c, in setting s, measured by UMS score;
- T_i is a binary indicator for the treatment assignment for individual i (1 if the student is assigned to treatment and 0 if not);
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (FE College, 6th Form College, School, Training Provider[16]), baseline attainment (measured through both KS2 raw mathematics scores and KS4 mathematics GCSE grade), and a dummy variable for whether a student was a recipient of Free School Meals;
- ε_{ics} is the idiosyncratic standard-error for individual i;
- β_o is the average intercept;
- γ_{cs} is the error for the class-level; and
- δ_s is the error for the setting-level.

For some settings, we only have one class of students; therefore we cannot estimate distinct class and setting level errors. However, if this proves detrimental to the estimation procedure as a whole, we will exclude estimating class-level errors. It must also be noted that for settings that have two teachers participating in the programme, there may be an additional clustering effect at the level of the teacher. Unfortunately, there are too few settings for which this is the case to include this as an additional component of the estimated error term.

Imbalance at baseline

We will assess imbalance at baseline, and for the sub-sample of those analysed, by calculating the following values in each case and cross-tabulating by treatment arm:

- Count and % male participants
- Count and % female participants
- Mean and standard deviation of KS2 maths attainment

[16] The category 'Training Provider' was not included in the trial protocol as it was not anticipated, but the developer has recruited a small number of these settings so we have added it here.

- Mean and standard deviation of KS4 maths GCSE grade
- Mean and standard deviation of setting type
- Count and % Ever 6 FSM

The difference in mean baseline attainment scores (KS2 scores and KS4 grades) between the treatment and control group will be expressed as Hedges' g effect sizes.

Missing data

First, we will report the number of complete observations (those without any data missing). This could identify the following types of missing data:

- Missing pre-treatment covariates
- Missing outcome data

Data can be missing completely at random (MCAR), missing at random (MAR), or missing not a random (MNAR). These data types and forms of missingness are summarised in the table and addressed in more detail below. In line with EEF guidelines on resource allocation, any imputation will be restricted to the primary analysis and will only be carried out when more than 5% of the data is missing. Schultz and Grimes (2002)[17] suggest that, when less than 5% of data is missing, there is likely to be little bias introduced to estimated treatment effects, so we have adopted this threshold here.[18]

Type of missing variable	Type of missing data	Further analysis
Pre-treatment covariates	MAR	Multiple imputation before analysis of treatment effects
Pre-treatment covariates	MNAR	Sensitivity analysis
Outcome data	NA	None. Missing outcome data will not be imputed and these observations will therefore be lost to analysis.

Missing pre-treatment covariates

All observations with missing pre-treatment covariates will be included in the analysis as long as the outcome measure and treatment assignment are not missing. It is likely that some observations will be missing values for baseline attainment (KS2 maths score and KS4 GCSE maths grade) due, for example, to pupils in the sample who did not sit the test(s). For any missing data of this type (where more than 5% of the data is missing), we will first try to establish which variables are predictive of the missing data. To do this, we will create a new variable that is a binary indicator of missingness and look for its predictors using a logistic regression model to establish correlations with other the variables in the dataset. Missing KS2 maths score data will be modelled as follows:

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

[17] Schulz, K. F., & Grimes, D. A. (2002). Sample size slippages in randomised trials: exclusions and the lost and wayward. *The Lancet*, *359*(9308), 784.
[18] This is also in line with the convention provided in EEF's guidance on statistical analysis.

where:

- M_{is} is the binary variable for missingness (equal to 1 if missing and 0 if not missing);
- p_{is} is the probability that a given observation is missing the KS2 maths score;
- x_1 is the EVER6_FSM_P variable;
- x_2 is the KS4 GCSE maths grade; and
- x_3 is the categorical variable for the setting type.

The same model will be used to model missingness for the KS4 GCSE maths grade, substituting KS4 GCSE maths grade for KS2 maths score in the specification above. Where both KS2 maths score and KS4 GCSE maths grade are missing, the following model will be used,

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

where:

- M_{is} is the binary variable for missingness (equal to 1 if missing and 0 if not missing);
- p_{is} is the probability that a given observation is missing both KS2 maths score and KS4 GCSE maths grade;
- x_1 is the EVER6_FSM_P variable; and
- x_3 is the categorical variable for the setting type[19].

A logistic regression model is appropriate here because the baseline probability of missingness is not known in advance, so whether a linear model would be a suitable approximation is also not known. Hence, using a logistic regression is more conservative (though the usual trade-off with interpreting the coefficients of logistic regressions will still apply).

If the coefficients in the regression are significant (i.e. the values are missing conditional upon other variables in the model) and missingness does not depend on unobserved covariates, imputation will provide an unbiased estimate of the true values. Multiple imputation (MI) will be carried out using the Markov chain Monte Carlo (MCMC)[20] method to predict the missing values prior to the analysis of treatment effects. We will then estimate the treatment effect using the imputed data and compare our result with the primary analysis (conducted on complete cases only).

If the missingness is correlated with cluster (as measured by a significance test) then we will substitute the above model with a three-level logistic regression with levels for classes and settings as in the HLM described in the robustness checks above. This multilevel model will then be used instead in the MCMC process for doing the multiple imputation.

If the point estimates of these two datasets are not similar (more than 0.05 standard deviations apart), it is likely that the data is MNAR and sensitivity analysis will be carried out. This will entail modelling missingness as above with all possible combinations of the

[19] Please note x_2 is excluded to keep consistency with the previous models specified above

[20] Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD*, *49*, 1-11.

variables available (EVER6_FSM_P, KS2 maths score and KS4 GCSE maths grade, and setting type).

If, after modelling missingness, as described above, it is found that our covariates do not explain the missingness, this will imply that the data is either MCAR or MNAR. In this case, we will be conservative and assume that the data is MNAR and conduct sensitivity analysis. These sensitivity analyses will investigate the sensitivity of the point estimate of the treatment effect to changes in model specification (and hence sample definition), through the inclusion and exclusion of variables for which observations are missing, as well as using null imputation to provide a more intuitive analysis based on a full sample of data.

Missing outcome data

No intermediate outcomes have been identified as appropriate to use to impute primary outcome data. Observations with missing primary outcome data will be dropped from the analysis and a complete case analysis will be run.

Compliance

In the context of the trial, we have defined compliance as the number of Maths-for-Life lessons a participating teacher delivers to students in each of the classes they teach, which can range between 0 and 5 for any given class (group). It must be noted that we are unable to collect student attendance data to each class delivered, which is a limitation with this definition of compliance. We will estimate the Complier Average Causal Effect (CACE) using a two stage least squares (2SLS) approach. We estimate:

$$T_{ics} = \gamma_0 + \gamma_1 . Z_i + \delta \mathbf{X}_{is} + u_{ics} \quad (1)$$

$$Y_{ics} = \beta_0 + \beta_1 . \widehat{T_i} + \alpha \mathbf{X}_{is} + \varepsilon_{ics} \quad (2)$$

where:

- Z_i is a binary indicator for the treatment (1 if the student is treated and 0 if not);
- T_{ics} is a continuous variable indicating the number of Maths-for-Life lessons a participating teacher delivered to individual i, in class c, in setting s;
- X_{is} is a vector of individual-level and setting-level covariates including a categorical variable for the type of setting (FE College, 6th Form College, School, Training Provider[21]), baseline Maths attainment (measured through both KS2 raw mathematics scores and KS4 mathematics GCSE grade), and controlling for whether a student was a recipient of Free School Meals;
- u_{ics} are Newey-West robust standard errors;
- ε_{ics} are Baum–Schaffer–Stillman 2SLS errors;
- $\widehat{T_i}$ are the predicted levels of compliance with the programme from (1); and
- Y_{ics} is the outcome for the KS5 mathematics GCSE resit performance for individual i, in class c, in setting s, measured by UMS score.

Intra-cluster correlations (ICCs)

ICCs will be calculated using a one-way analysis-of-variance (ANOVA) model. Given randomisation is at the setting level, we will only calculate the ICC at this level. This will only

[21] The category 'Training Provider' was not included in the trial protocol as it was not anticipated, but the developer has recruited a small number of these settings so we have added it here.

be calculated post-test as we are not collecting a pre-test primary outcome, rather using KS4 grades collected from the NPD. Additionally, it would not be valid to calculate ICCs pre-test given the trial is conducted in FE institutions, and many of the students have moved institution between KS4 and KS5.

Effect size calculation

Effect sizes will be expressed in terms of Hedges' g, using the following formula:

$$\boldsymbol{ES} = \frac{\boldsymbol{M_1} - \boldsymbol{M_2}}{\boldsymbol{SD^*_{pooled}}}$$

Where,

- $\boldsymbol{M_1}$ is the mean value of the outcome in the control group; and
- $\boldsymbol{M_2}$ is the mean value of the outcome in the treatment group.

And,

$$\boldsymbol{SD^*_{pooled}} = \sqrt{\frac{(\boldsymbol{n_1} - \mathbf{1})\boldsymbol{SD_1^2} + (\boldsymbol{n_2} - \mathbf{1})\boldsymbol{SD_2^2}}{\boldsymbol{n_1} + \boldsymbol{n_2} - \mathbf{2}}}$$

Where,

- $\boldsymbol{n_1}$ is the effective number of observations analysed in the control group, accounting for the design effect due to clustering;
- $\boldsymbol{n_2}$ is the effective number of observations analysed in the treatment group, accounting for the design effect due to clustering;
- $\boldsymbol{SD_1}$ is the standard deviation of the outcome variable in the control group; and
- $\boldsymbol{SD_2}$ is the standard deviation of the outcome variable in the treatment group.

The difference in means for each outcome variable will be recovered from the relevant regression equation above ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$).

Appendices

Appendix 1: Analysis code

In this appendix, we provide indicative analysis syntax to implement the models specified in the Statistical Analysis Plan using Stata 14. 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.

Primary outcome analysis

```
regress ums assigned type ks2maths ks4maths everfsm_6_p, vce(cluster
setting)
```

is a linear regression model estimated on the full randomised sample data where `ums` is the UMS score for the individual's GCSE maths resit exam (corresponding to Y in the regression equation), `assigned` is the treatment assignment, `type` is the setting type, `ks2maths` is the KS2 maths score of the individual, `ks4maths` is the KS4 maths GCSE grade of the individual and `setting` is specific setting in which the individual is clustered.

Secondary outcome analysis

The same syntax will be used for the secondary analysis as the primary analysis, replacing the variable names as appropriate.

Subgroup analysis

```
regress ums assigned type ks2maths ks4maths assigned#everfsm_6_p,
vce(cluster setting)
```

is a linear regression model estimated on the full randomised sample data where `ever6_fsm_p` is the FSM status of the individual and all other variables are as above.

Robustness checks

```
xtmixed ums assigned type ks2maths ks4maths everfsm_6_p, || class:, var ||
setting:, var
```

is a hierarchical linear model estimated on the full randomised sample data where `ums` is the UMS score for the individual's GCSE maths resit exam (corresponding to Y in the regression equation), `assigned` is the treatment assignment, `type` is the setting type, `ks2maths` is the KS2 maths score of the individual, `ks4maths` is the KS4 maths GCSE grade of the individual, `class:` allows the intercept to vary across classes, and `setting:` allows the intercept to vary across settings.

CACE analysis

```
ivregress 2sls ums comply type ks2maths ks4maths (comply = assigned),
vce(cluster setting)
```

is an instrumental variables (two stage least squares) regression model estimated on the full randomised sample data where `comply` is a continuous indicator of compliance and all other variables are as above.