Statistical Analysis Plan for 'Texting Students and Study Supporters' ('Project SUCCESS')

NatCen

PROJECT TITLE	Evaluation of the Texting Students and Study Supporters Programme ('Project SUCCESS')
DEVELOPER	BIT
EVALUATOR	NatCen Social Research
PRINCIPAL INVESTIGATOR	Julia Griggs
TRIAL (CHIEF) STATISTICIAN	Neil Smith
SAP AUTHOR(S)	Neil Smith, Karl Ashworth
TRIAL REGISTRATION NUMBER	ISRCTN70011940
EVALUATION PROTOCOL URL	(see hyperlink for amended protocol)

VERSION	DATE	REASON FOR REVISION
1.1		
1.0	14/05/2018	N/A

Table of Contents

Introduction

Project Success is an intervention for students resitting GCSE English and/or Maths developed by the Behavioural Insights Team (BIT). The programme aims to improve students' study habits, build students' peer and adult support networks, improve student attendance and ultimately improve GCSE attainment.

Previous research indicates that texting parents can have a positive impact on pupils' attainment and attendance. Two recent trials have found that text messages sent to a nominated study supporter increased further education (FE) college students' attendance[1]. A similar cluster randomised controlled trial (RCT) of texting parents ahead of an upcoming test had a positive impact on maths performance and attendance[2]. Furthermore, another RCT found that children whose parents received weekly texts from tutors concerning specific issues were less likely to be absent or drop out[3].

The intervention is being delivered across 31 post-16 study institutions[4] and consists of 36/37 text messages[5] sent to the student and/or a named study supporter over the course of one academic year. The text messages provide information on course content, course deadlines, available additional resources, tutorial sessions and examination timetables.

Different approaches to the programme are being tested: 1) texts targeted at the student, 2) texts targeted at a study supporter, e.g. a parent or peer identified by the student, or 3) texts targeted at both. The quantitative analysis will estimate the effect size of each intervention compared to the single "business-as-usual" control group. The primary outcome will be the percentage of students that pass their GCSE mathematics or English resit exams after one year (the subject they are receiving texts about). Students' attendance will be considered as a secondary outcome measure.

Study design

Population and Eligibility Criteria

The intervention is being delivered in further education (FE) colleges in England. FE colleges were selected for the trial because they contain the highest proportion of post-16 students resitting GCSE English and/or maths.

Young people are eligible to take part in the trial if they are:

- Aged 16 or older
- Enrolled at one of the participating colleges in September 2017
- Due to resit GCSE maths and/or English in the 2017/18 academic year.

The developer (BIT) used a targeted recruitment strategy to engage FE colleges with the intervention. An initial screening call was used to gauge colleges' interest and ascertain important information, such as the number of students eligible for the trial. A Memorandum of Understanding (MOU) was sent to interested colleges to sign. The MOU included an overview of the intervention and details

[1] Groot et al. (2017) 'I get by with a little help from my friends Two field experiments on social support and attendance in further education colleges in the UK', http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2017/04/Study-Supporter-WP_April-2017.pdf [Accessed 02/11/2017]
[2] Miller et al. (2016) *Texting Parents: Evaluation report and executive summary,* available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Texting_Parents.pdf [Accessed 02/11/2017]
[3] Kraft, M. A. and Rogers, T. (2015) 'The underutilized potential of teacher-to-parent communications: Evidence from a field experiment', *Economics of Education Review,* 47, pp. 49-63.
[4] The total number of colleges deviates from the trial protocol (n=32). One college, which had previously been counted as an independent unit, has since been identified as being part of a wider group (i.e. a separate campus rather than separate college).
[5] Maths students have one additional exam, and therefore receive an additional reminder text.

about the requirements for participating colleges. College leaders were asked to give their consent and return a signed copy of the MOU to the developer.

Students were recruited to the intervention in September and October 2017 using an online survey provided to college 'project leads' by the Behavioural Insights Team. The survey included information about 'Texting Students and Study Supporters' ('Project SUCCESS') and what would be expected of participants, so that they were able to give informed consent. The survey also included a small number of questions about students' attitudes towards learning and their motivation to study. Students willing to take part were asked to nominate a named study supporter, describe their relationship and provide their study supporters' contact details. Study supporters were not asked to give their consent to take part in the trial, but were able to opt-out of the intervention when the first text message was sent (in early November 2017).

Trial Design/ Description of trial arms

Project SUCCESS is being evaluated using a four-armed cluster randomised controlled trial (RCT) with randomisation to each of the four trial arms at the individual level. Clustering within colleges will be accounted for as a fixed effect in the model. The four arms of the trial are:

- Student receives text messages
- Study supporter receives text messages
- Student and study supporter receive text messages
- Control (no text messages)

Sample size

The trial is being conducted in 31 FE colleges.[6] An average of 123 eligible students per college have given their informed consent to participate in the trial; a total of 3,823 individuals. Students retaking both maths and English have been assigned at random to either the maths or English strand of the intervention, so that approximately 50 per cent are participating in each stand.

Number and timing of measurement points

There will be a single post-intervention measurement point. GCSE resit attainment in maths and English post-treatment will be taken from the National Pupil Database (NPD) in autumn 2018 for all students who participated in the programme in the 2017/18 academic year. Students retaking both English and maths will be randomly assigned to receive texts about just one of these subjects and the primary outcome will be attainment in the subject the student received texts about.

Randomisation

A multi-site randomisation design was used, with the FE college as the stratifying variable. Students were randomly allocated on a 1:1:1:1 basis to one of four intervention groups (three treatment and one control). It had been agreed at the outset that students retaking both English and maths would receive texts about one subject only. Therefore those students retaking both were randomly allocated to either one or the other outcome group to ensure an overall balanced design, giving an overall average of 31 students per trial arm per college (15/16 per arm, per subject), i.e. an average of 123 students per FE college.

Focusing the design on one subject per person from those taking dual re-sits protects the analysis from potential spill-over effects. For example, a person receiving an English text treatment reminder might also act upon that reminder to do further work on their maths. If that person's maths exam results were included in the non-texting control group this spill-over reminder effect could bias the impact estimate. By allocating dual resit students either to maths or English and removing their results

[6] Please see footnote 4 for an explanation of how and why this number deviates from the trial protocol.

on the unallocated subject from the analysis, any spill-over effects on the unallocated subject will not be included in the analysis and so cannot influence the results.

Randomisation to each of the four treatment groups was carried out in Stata MP Version 14, using the `.randomise` command with a maximum of 100 randomisations used to achieve balance on the subject taken. The randomisation was conducted at the end of October 2017 by an independent analyst within the evaluation team. Table 1 provides a breakdown of numbers randomised by college and trial arm.

Table 1: Results of the randomisation

College[7]	Arm 1	Arm 2	Arm 3	Arm 4	Total English	Total maths	Total per college
College 1	51	51	52	51	104	101	205
College 2	8	8	8	9	22	11	33
College 3	18	18	17	18	49	22	71
College 4	8	9	9	9	15	20	35
College 5	42	43	43	43	84	87	171
College 6	22	22	22	22	45	43	88
College 7	26	26	26	25	67	36	103
College 8	33	33	33	32	44	87	131
College 9	23	23	23	24	80	13	93
College 10	29	28	29	28	53	61	114
College 11	33	32	32	32	51	78	129
College 12	6	6	6	6	14	10	24
College 13	50	49	49	49	77	120	197
College 14	27	28	27	27	19	90	109
College 15	22	22	22	22	58	30	88
College 16	51	51	51	52	147	58	205
College 17	13	13	13	13	19	33	52
College 18	16	16	16	15	27	36	63
College 19	30	30	30	30	50	70	120
College 20	14	14	15	15	27	31	58
College 21	46	46	46	47	48	137	185
College 22	43	42	42	43	82	88	170
College 23	45	45	45	44	117	62	179
College 24	25	25	25	25	55	45	100
College 25	64	64	63	64	102	153	255
College 26	18	18	18	17	53	18	71
College 27	12	13	12	13	22	28	50
College 28	21	21	21	21	34	50	84
College 29	73	73	73	73	138	154	292
College 30	71	70	70	71	169	113	282
College 31	17	16	16	17	38	23	61
Total	957	955	954	957	1,912	1,911	3,823

[7] A further three colleges were initially recruited by BIT, but decided to leave the trial before randomisation.

Calculation of sample size

The primary study outcome has been defined as a binary variable – whether or not the student achieves a pass in their maths or English GCSE resit examination. EEF guidance recommends the use of the relative risk ratio as an effect size measure for binary outcomes.[8] The Relative Risk Ratio (RR) can be calculated from the log-odds ratio, which itself is parameterized within a logistic regression model; which, in turn, is appropriate for the analysis of binary outcomes, given that it bounds predicted probabilities between zero and unity.

Arguably, the sample size calculation would be based upon the effect size parameterised through the logistic model framework. Unfortunately, as noted by Demidenko[9] (2007) '*there is no consensus on the approach to compute the power and sample size with logistic regression'*. Moreover, McConnell & Vera-Hernandez[10] (2015), following Schochet[11] (2013), argue that modelling differences in probabilities is more intuitive. Given the lack of consensus in dealing with power analysis for binary outcomes and, particularly, the lack of an agreed approach for complex logistic models with clustered designs, we have followed standard practice for differences in outcome probabilities, using the binomial variance [p * (1-p)] as the plug-in (e.g. McConell and Vera-Hernandez, 2015). From Maynard and Dong[12] (2013, p. 50), we have the following formula:

$$MDES = M_{J-g^*-1}\sqrt{\frac{\rho\omega(1-R^2_{2T})}{J} + \frac{(1-\rho)(1-R^2_1)}{P(1-P)Jn}} \qquad (1)$$

The MDES is the standardised minimum detectable effect size, i.e. the minimum impact we would expect the experiment to detect with the given sample size and design. Where M_{J-g^*-1} is the sum of the t values corresponding to the appropriate levels of the Types I and II error levels, with the appropriate degrees of freedom, R^2 is the proportion of variance explained by covariates entered at the appropriate level of the hierarchical model; ρ is the intra-cluster correlation, ω is the between site treatment heterogeneity effect, J is the number of clusters (colleges) and n is the average number of students per college.

Given (from Bloom, 2006;[13] equation 4) that

$$ES = \frac{\bar{Y}_t - \bar{Y}_c}{\sigma}$$

We can simply multiply the MDES, calculated through PowerUp, by σ to get the MDE, i.e. the minimum impact effect stated as a difference using the scale of a proportion. We know σ from the binomial nature of the data, i.e. $\sigma^2 = \theta \times (1-\theta)$, where we will take θ to be the proportion anticipated achieving success in the control group.

Our assumptions in calculating the MDE are as follows:

- We have a 2-level multi-level model with allocation to treatment modelled as a fixed effect at level 1 (students), with colleges (level 2) allowed to vary randomly with respect to student outcomes (i.e. $r_{ij} \sim N(0, \sigma^2)$.

[8] Statistical Analysis Guidance, 2018, p. 5.
[9] Demidenko, E. (2007) Sample Size Determination for Logistic Regression Revisited. Statistics in Medicine, 26, pp. 3385-3397, available at: http://dx.doi.org/10.1002/sim.2771
[10] Going beyond simple sample size calculations: a practitioner's guide https://www.ifs.org.uk/publications/7844
[11] Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference http://journals.sagepub.com/doi/pdf/10.3102/1076998611432176
[12] Dong, N. and Maynard, R. A. (2013) PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. Journal of Research on Educational Effectiveness, 6(1), pp. 24-67, doi: 10.1080/19345747.2012.673143
[13] The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology, available at: https://files.eric.ed.gov/fulltext/ED493363.pdf

- We assume a total of 31 participating colleges and an average of 60 students per college for each of the three pairwise comparisons, with 30 in each treatment arm and 30 in the control (with a broadly even split between maths and English). This means a total of 930 students in each treatment arm and 930 in the control group; a total of 1860 for each of the three pairwise comparisons.
- Maths and English outcomes (i.e. a pass or fail in the English or maths GCSE resit exam) will be considered together as a single outcome variable.
- We assume equal proportions of students in the treatment and control groups (P = 0.5) for each pairwise comparison.
- We anticipate a baseline pass rate of 30% ($\theta = 0.3$)
- We have three separate significance tests for the primary outcome, with each of the three treatment groups separately compared to the control group.
- We assume a nominal Type 1 error rate of 0.05 with a two-tailed test, amended using a Bonferroni adjustment for three separate independent tests, to a value of 0.0167 for each test. The Type II error rate is set at 0.2, giving power of 0.8.
- We assume[14] an intra-cluster correlation coefficient (ρ) of 0.1.
- No between site heterogeneity is expected (i.e. $\omega = 0$).
- No level 1 baseline test covariate has been included in the calculation, for reasons explained below. There are also no level 2 stratification variables because we have treated the colleges as random effects and no preliminary stratification was undertaken for college type.
- We assume no attrition in the design, our rationale for which is explained below.
- The EverFSM MDE calculation combined all three treatment arms and assumed a 20% prevalence per college, giving (90 x 0.2) + (30 * 0.2) = 24 students per college.

The assumptions for each MDE estimation are given in Table 2, below.

The sample size calculations in Table 2 were originally intended to enable the detection of a seven percentage point increase in the GCSE pass rate, from (an estimated) 30 per cent to 37 per cent.[15] However, the underlying model for the purposes of this power calculation, which we have now changed as described above, assumed, a fixed effects model, with no random variation at Level 2, i.e. colleges represented strata from a population fully defined by the sample.

Table 2 presents, using the current model assumptions, the intention-to-treat minimum detectable effects (MDEs) - defined as the minimum detectable difference in the probability of passing GCSEs between the treatment group and the control group. Since conducting the original calculation, we have amended the assumptions from a fixed to a random effects model and added one further college.[16] The estimated MDE is now plus/minus 6.8 percentage points, so is consistent with the previous estimate of 7 percentage points.

We have assumed a baseline control group pass rate of 30 per cent. However, no consistent baseline test is available to help reduce sample size requirements; subsequently our power calculations ignore

[14] It is worth noting that from (1) when we set $\omega = 0$ and with both R^2 values set to zero, as ρ reduces to zero, the precision decreases, i.e. in contrast to a randomised cluster design a higher ρ is beneficial to the precision of the estimator. When ρ=0 formula 1 equates to a single level design calculation, with appropriate adjustments to the degrees of freedom in the t-value.

[15] A 10% difference is achievable with a low level resource, whereas power calculations to reach a 5% level of difference suggest considerably more intensive resource requirements. Therefore, the compromise of 7% was agreed, giving a realistic and achievable measure of detectable difference which can be measured within the logistical and budgetary constraints.

[16] Original estimates were based on 30 participating colleges. However, the developers (BIT) were able to recruit an additional FE college. It is important to note three of the colleges within this total are part of a wider educational organisation, and therefore represented by one project lead.

this parameter.[17] We also provide an MDE (10.7 per cent) for young people who receive or have received free school meals (EverFSM), using the 30 per cent baseline.

Table 2: Minimum detectable effects – intention-to-treat analysis[18]

	Treatment A vs. Control	Treatment B vs. Control	Treatment C vs. Control	FSM: Treatment combined vs. Control
Alpha	.0167	.0167	.0167	.05
Power	.8	.8	.8	.8
Number of colleges	31	31	31	31
Students per college	60	60	60	24
Proportion of sample in treatment group	0.5	0.5	0.5	0.75
Intra-cluster correlation	0.1	0.1	0.1	0.1
MDE	6.8	6.8	6.8	10.7

Notes: There is no treatment heterogeneity, no covariates entered at either Level 1 or Level 2, the baseline proportion successful is assumed to be 0.3, no attrition – attrition values are assumed to be coded zero. We have assumed a sample size for EverFSM of 0.2 of overall sample size. Treatments A, B and C refer to the situations where A) only the student receives a text, B) only the supporter receives a text, and C) both student and supporter receive the text.

Attrition and Follow-up

Individual level attrition may occur through students formally withdrawing their consent to receive texts or by changing their phone number and not informing the experimental administrators of their new number.[19] Students may also drop-out of college or fail to re-sit their exams. We do not envisage any college level attrition at this point; given that their role in the design is minimal and requires little effort. However, where students withdraw from the experiment either through formally requesting no further texts or by failing to update a changed phone number, we still have their outcome data from the NPD, provided they resit their exam, and we propose to use this in the analysis.[20]

The design envisages no formal follow-up even though pupils may decide not to take one or both of their resits, or may leave college prior to taking their re-sits. In such cases, we will assign their outcomes to a fail; this is justified because the policy interest is in students passing the exam rather than the reasons for failure (i.e. we make no distinction through not taking and performing poorly in the exam). This procedure helps to avoid potential selection effects that could arise from removing

[17] It was anticipated that BKSB/ ForSkills scores could be used as a measure of baseline academic achievement (as set out on page 7 of the protocol); however, further discussion with colleges suggests that this data will only be available in a very small minority of cases, and thus will not provide a suitable measure. Using KS4 (GCSE) scores was also discussed. However, it was felt that as the new requirement to retake GCSE maths and English specifically targets students who received a grade of D/level 3, baseline GCSE grades were likely to vary very little among the sample, and therefore offer little value. It is not possible to obtain GCSE raw scores (i.e. percentages) via the NPD.

[18] MDE calculations based on formulas for the calculation of sample sizes for binary outcomes outlined in McConnell and Vera-Hernández (2015) Going Beyond Simple Sample Size Calculations: a Practitioner's Guide. IFS Working Paper W15/17. Institute of Fiscal Studies. Available at: https://www.ifs.org.uk/uploads/publications/wps/WP201517.pdf

[19] The BIT texting platform allows the developer to monitor the number of 'bounce-backs' (i.e. messages which could not be delivered). This number is available for each weekly 'batch' of text messages sent.

[20] It is possible that students may request their data not be used in the experiment. If this happens we will consider treating them as missing data and, if appropriate, apply the weighting strategy described for dealing with missing data.

such events from the primary analysis. Consequently, all students starting the experiment will be included in the analysis under the intention-to-treat estimator.

In addition, students who used their own telephone number when asked to provide contact details for a study supporter[21] will be included in the group into which they were allocated for treatment. We will address non-compliance through the calculation of causal average complier estimates (see below).

Outcome measures

Primary outcome

Primary outcome: GCSE resit attainment in maths or English post-treatment will be taken from the National Pupil Database (NPD) in autumn 2018 for all students who participated in the programme in the 2017/18 academic year. For those students taking both English and maths we will include only the subject relevant to the group to which they were allocated, i.e. English or maths, as appropriate. We will pool into a single group all students allocated to the maths group with all students allocated to the English group so that the success rate is defined as a pass in English or a pass in Maths. This gives us an average of 30 students in each treatment arm per college: 15 in English and 15 in maths, combined into a single group of 30 students. Consequently, each of the three pairwise analyses will comprise 60 students per college, i.e. 30 in the treatment group and 30 in the control group. Overall, across all three treatments and the control group there are 60 x 4 = 320 students on average per college in the experiment.

Analysis of the free school meal (FSM) students will combine all three treatment groups into a single group and compare them to the control group (i.e. a ratio of 3:1). We assume that 20 per cent of the overall group will be EverFSM, i.e. 48 out of 240 in each college.

We understand that the primary concern of the study is whether or not students pass their GCSE resit. Consequently, a binomial pass/fail measure is more appropriate than a continuous outcome for the current study. In addition, it is anticipated that re-sit scores will cluster around the pass/fail threshold, leaving a restricted range of scores on the Level 1-9 scale, which will not be suitable for analysis assuming a continuous measurement scale.[22] (Unfortunately, it is not possible to obtain the raw GCSE marks (i.e. a percentages) underlying these grades from the NPD.)

In view of this, we propose that English and maths attainment scores for the end of the academic year of study be taken from the NPD and coded in binary format: a KS4 re-sit pass is defined as attaining level four or above; a fail is defined as reaching levels one to three. We will also classify young people who do not re-sit the exam as a 'fail', in line with the intention-to-treat approach. Outcome data will be obtained by matching pupils at the start of the academic year in the study to their outcome data for the end of the academic year. The outcome estimate will be a pass-rate proportion (a ratio of the number who started the trial and passed the exam divided by number starting the trial) bounded by zero and unity.

Each outcome has four intervention groups: three treatment and one control, and the aim is to test each treatment separately against the control group to determine the efficacy of each treatment. Consequently, we run three paired treatment–control tests. In order to maintain a five per cent significance level for each test, a standard Bonferroni adjustment was applied to account for the three tests in each of the families. This resulted in a value of 0.0167, i.e. α/3, and only tests showing a significance level below this threshold will be accepted as statistically significant at alpha = 0.05.

[21] Analysis suggests that around 250 students entered their own number rather than their study supporter's, but that these cases were evenly distributed across colleges and treatment group. This issue will be discussed in more detail in the final evaluation report.

[22] The NPD does contain semi-continuous scores (from 1-9) for GCSE maths and English resit students. However, students retaking exams as part of the new requirement are likely to be at a very similar academic level, thus likely to obtain results which cluster around the same two or three grades (3-5). We therefore feel there is very limited value in using these grades.

Secondary outcomes

Students' records of attendance will be collected from colleges directly using a secure data transfer; (datasets being shared via secure Natcen's File Transfer Protocol (FTP) site). This will be a simple attendance record stating whether students were present or absent for each class. This will allow for the creation of a variable recording the overall attendance rate, i.e. the proportion of classes attended to all classes given for the course (i.e. attendance at each class taken as a binary variable), for each pupil in the study.

Data on student aspirations, confidence and attitudes to learning will be recorded as part of the process evaluation and does not constitute part of the impact evaluation.

Analysis

Primary intention-to-treat (ITT) analysis

It is proposed to run an intention-to-treat analysis, with a binomial outcome defined as passes in the numerator and failures and no shows in the denominator, as described above. This will involve running separate models for each of the three paired outcomes, i.e. each treatment group compared separately to the control group, where the significance level of the test will be conducted using the adjusted Bonferroni level. We propose to use a binomial multilevel logistic regression model, with students at Level 1 nested within colleges at Level 2, and colleges will be allowed to vary randomly, with a residual mean centred at zero with a Gaussian distribution.

The basic form of the model is:

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \quad (2)$$

We will run separate models for each treatment comparison group as follows:

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Student\ text\ only)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Supporter\ text\ only)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Student\ and\ Supporter\ text\ jointly)$$

Where the outcome variable is regressed onto the treatment binary indicator. All analyses will be conducted using the Stata analysis package.

The analysis does not adjust for academic achievement at baseline due to the lack of available data (see footnote 17).

Imbalance at baseline for analysed groups

Randomisation, if conducted correctly, should result in equivalent distributions of characteristics in the treatment and control groups. Any such differences arising will do so by chance. We will explore the potential for chance imbalances first through an inspection of the descriptive statistics of various characteristics, comparing treatment and control groups. Secondly, we will run a statistical model to identify any characteristics associated with any imbalance.

The NPD data received as standard in an NPD extract (i.e. data from Tiers 1 and 2, e.g. age, gender, everFSM) can be used to predict imbalances in the treatment and control group through regressing the treatment/control binary indicator onto the suite of extracted NPD variables. We propose to explore this using the model form (2) above with the logit transform, as described in the text. Any coefficients found to be significant will indicate imbalance and these variables will be included in an extension of model (2) above as a sensitivity test on the impact effect; i.e. we will extend model (2)

with significant predictors from the imbalance model and compare the magnitude of the impact estimate between the two impact estimate models.

Missing data

Based on past experience we anticipate missingness to be relatively low for the outcome variables. Missingness is of concern primarily if the variables with missing data are (i) related to the outcome and (ii) the pattern of missingness differs between treatment and control groups. If variables with missing values are unrelated to outcomes then there is no need to adjust for them. Such an adjustment will not reduce bias but will likely decrease precision through increasing the variance of the impact estimator, either through weighting or imputation variance adjustments.

If variables with missingness relate to outcomes, but the pattern of missingness is the same between treatment and control groups, again there is no need to adjust for missingness. The key concern is whether missingness is related to outcomes and whether the pattern of missingness differs between the treatment and control groups.

Consequently, to adjust for any missingness in the outcome variables we want first to establish whether variables used to predict missingness are associated with the outcome variable. Once we have identified relevant variables with missingness the second step is to assess whether the pattern of missingness differs between treatment and control groups.

For the first step we will identify the association between our suite of variables in the NPD extract and our outcomes. Secondly, variables that are correlated with outcomes will be entered as predictors[23] into a logistic regression model where the dependent variable is the missing/non-missing indicator on the outcome variable. This will inform us of which predictor variables are related to missingness in the outcomes. The next step will be to enter the treatment indicator and a series of interaction terms between the treatment indicator and those variables with a significant main effect association with missingness. Finally, assuming significant associations are found for some of the interaction terms, we will run a reduced form of the model including only significant interaction terms and their associated main effects from the previous model. From this we will construct a missingness weight as the inverse of the propensity score from the reduced model. We will then rerun models (1) and (2) above to explore the adjustment on the magnitude of the impact estimator. Stata's complex survey suite of models will be used (svyset) in order to adjust appropriately for the weight on the variance estimator of the impact effect.

Non-compliance with intervention

There is no a-priori clear cut measure of the number of texts a student should receive for the treatment to be effective; consequently, we do not have a theoretical basis to assign compliance as a function of the number of texts sent and/or received. However, we can consider the proportion defined as a function of the number of texts received divided by the maximum number of texts issued to an individual. In the first instance we will consider compliance as receiving all texts and non-compliance as anything less than all. Depending upon the distribution of texts received, we will consider lowering the threshold for compliance to less than 100 per cent to a lower threshold and using values above that threshold as compliance and below as non-compliance.[24] This will act as a sensitivity test for our compliance estimate.

[23] We intend to deal with missingness in predictor variables through dummy coding missing as an explicit categorical indicator for each variable included in the model.
[24] The BIT texting platform allows the developer to monitor the number of 'bounce-backs' (i.e. messages which could not be delivered). This number is available for each weekly 'batch' of text messages sent.

A common approach to estimating compliance has arisen from Bloom (1984)[25] and Imbens and Angrist (1994)[26] which use an instrumental variables approach to identify compliance. Allocation to treatment group under randomisation acts as an instrument allowing estimation of treatment on those who have complied with treatment. This approach identifies four possible groups of people in terms of their potential assignment and response to treatment offer:

- Defiers: people who do the opposite to whatever action they are assigned
- Never takers: people who always refuse treatment on offer
- Always takers: people who find treatment whatever their allocation
- Compliers: people who take the treatment if offered, but otherwise do not

We assume that never-takers and always-takers do not contribute to the treatment effect (e.g. Bloom, 2006; p10), leaving a treatment effect created by compliers and defiers. Thus:

$$LATE = \frac{ITT}{E(D|Z=1)-E(D|Z=0)} \quad (3)$$

Where LATE is the local average treatment effect and represents the treatment effect as the impact on compliers. Z refers to the instrument, i.e. the group to which a person is allocated through random assignment and D represents treatment received (0 = no, 1 = yes).

In the present design we consider it extremely unlikely that people outside of the treatment group seek out supportive texts in a way that mirrors treatment.[27] Consequently, we assume that $E(D|Z=0) = 0$. Formula 3 then reduces to:

$$TOT = \frac{ITT}{E(D|Z=1)} \quad (4)$$

Where TOT is treatment on the treated. Typically, the TOT represents a weighted average of always takers and compliers.

We note that we can represent the ITT as an odds ratio or on proportional scale, i.e. the difference between two proportions, and will present both versions in the report, in compliance with EEF guidelines for binomial outcome data. We will calculate ITT and TOT on both scales for all impact analysis undertaken.

We also note that potentially we can extend the compliance estimator using the distribution of the proportion of compliance to create separate 'groups' for each level of compliance (e.g. group 1 = compliance < 25%, 2 = 50% etc.). If we have sufficient numbers in each group, we could calculate an 'average causal response with treatment intensity' (ACLTI) (Angrist and Pischke, 2009, Section 4.5.3).[28] This approach uses two-stage least squares to weight together the estimates across the compliance groups. We will update the analysis plan accordingly should the distribution of non-compliance[29] allow for a reasonable estimation of ACLTI.

Secondary outcome analyses

We originally proposed in the protocol to analyse attendance in class as a continuous outcome defined as the proportion of classes attended out of the total number taught. On reflection, it seems

[25] Accounting for No-Shows in Experimental Evaluation Designs https://doi.org/10.1177/0193841X8400800205

[26] Identification and Estimation of Local Average Treatment Effects http://www.nber.org/papers/t0118

[27] In principle, it is possible that people send themselves weekly text reminders which are similar to those used in the experiment but we will not know that from our design.

[28] Angrist, J. D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

[29] We will not know the full extent of non-compliance, i.e. drop out/bounce-back until the end of the experiment because people can drop out at any time.

more appropriate to consider this as a binomial outcome because the binomial distribution will bound predicted proportions between unity and zero. In other words, attendance at each session will be treated as a binary outcome (attended/not attended). We will use a logit model to explore the proportion of sessions attended versus the total number of sessions. In addition, we will amend the proposed CACE analysis to be appropriate to a binary outcome (see below). Therefore, analysis of the secondary outcomes takes the same form as model 2, where $\pi_i = {}^{Y_i}/_{X_i}$, where X is the total number of classes and Y is the total number of classes attended, i.e.

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Student\ text\ only)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Supporter\ text\ only)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Treatment = Student\ and\ Supporter\ text\ jointly)$$

Subgroup analyses

We will approach the analysis for students who have ever been eligible for Free School Meals (EverFSM) in two ways. First, we will combine all treatment groups into a single treatment group and then subset to include only those who have ever had free school meals, as recorded on the NPD (EverFSM). Secondly, Model (2) will be extended through the inclusion of a dummy variable for FSM recipients (variable = EverFSM) and a corresponding interaction effect between FSM and treatment to test for a differential impact on FSM recipients. The specifications are:

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1j} treatment_{ij} + r_j \ (Domain = EverFSM, Treatment = All\ groups\ combined)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1ij} treatment_{ij} + \beta_{2ij} EverFSM_{ij} + \beta_{3ij} EverFSM.treatment_{ij} + r_j \ (Treatment = All\ groups\ combined)$$

As proposed in the protocol, we will also explore whether there is evidence for a differential effect on maths and English outcomes, gender and sitting either English or maths versus sitting both English and maths:

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1ij} treatment_{ij} + \beta_{2ij} Subject_{ij} + \beta_{3ij} Subject.treatment_{ij} + r_j \ (Treatment = All\ groups\ combined)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1ij} treatment_{ij} + \beta_{2ij} Gender_{ij} + \beta_{3ij} Gender.treatment_{ij} + r_j \ (Treatment = All\ groups\ combined)$$

$$logit(\pi)_{ij} = \beta_{0j} + \beta_{1ij} treatment_{ij} + \beta_{2ij} N\_Subjects_{ij} + \beta_{3ij} N\_Subjects.treatment_{ij} + r_j \ (Treatment = All\ groups\ combined)$$

Effect size calculation

The EEF guidance suggests the use of a relative risk ratio (RR) as an effect size for binary outcomes. Consequently, this is what we propose to use here for each of the three models making up our primary outcomes. We note that the primary outcome models proposed here are relatively simple and do not include controls for covariates. This is advantageous given that the RR can vary conditional upon covariate levels and would require the centring of covariates in the logistic model to assess the RR at its average covariate value. However, we do not have this issue with the models proposed here, but the Level 2 random college effect will translate into the odds for students at the average level of success across colleges.

The protocol did not specify the calculation of effect sizes for the primary outcomes but did propose an effect size appropriate to continuous outcomes for class attendance. As noted above, we now propose to treat the model for class attendance using a binomial distribution appropriate to modelling differences in proportions. We therefore propose to use the relative risk ratio as a measure of the effect size for this outcome

We will calculate the relative risk ratio, following a method outlined by Fleiss & Berlin (2009),[30] as the measure of effect size, using the results, as appropriate, from the three versions of model (2) above for each of the treatment groups and for the class attendance model.

$$RR = \frac{1+exp^{-\alpha}}{1+exp^{-\alpha-\beta}} \quad (5)$$

In (5) above α refers to the constant from the logistic model (2) and β refers to the treatment indicator.

.

[30] Cooper, H., Hedges, L. and Valentine, J. *The handbook of research synthesis and meta-analysis*