BITUP: Updating Parents on Number of School Days Missed, a two-armed cluster randomised trial
Statistical Analysis Plan

Evaluator (institution): Verian
Principal investigator(s): Prof Natalie Gold

PROJECT TITLE	BITUP: Updating Parents on Number of School Days Missed, a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	The Behavioural Insights Team (trading as Behavioural Insights Ltd)
EVALUATOR (INSTITUTION)	Verian (trading as Verian Group UK Ltd)[1]
PRINCIPAL INVESTIGATOR(S)	Prof. Natalie Gold
PROTOCOL AUTHOR(S)	Pieter Cornel, Dr Sarah Bowen, Dr Debbie Blair, Dr Michael Ratajczak, Ben Toombs, Prof. Natalie Gold
TRIAL DESIGN	Randomised controlled trial with family-level randomisation
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	11-16, KS3-4
NUMBER OF SCHOOLS	108
NUMBER OF PUPILS	104,029
PRIMARY OUTCOME MEASURE AND SOURCE	Pupil absence rate (school data via Wonde)
SECONDARY OUTCOME MEASURE AND SOURCE	Pupil unauthorised absence rate and pupil authorised absence rate (all from school data via Wonde)

SAP version history

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

[1] The Evaluators were previously known as Kantar Public UK, and their name was changed to Verian in November 2023.

Table of contents

Contents

Introduction

This statistical analysis plan sets out the intended impact evaluation for the independent impact evaluation of the BITUP intervention in 108 schools across England. The BITUP intervention aims to improve pupil attendance among pupils in Year 7 – Year 11 in secondary schools by improving parental awareness of their child(ren)'s attendance in the previous term (5-8 week period) and their understanding of the impact of absenteeism on educational outcomes. The intervention is delivered through schools' usual text messaging procedures at the start of each new term, to the primary parent or carer phone number for each pupil of each pupil in the intervention group with attendance <95% in the previous term.

Content of messages: The two core components of the intervention are:

1) Simplification: messages contain the **number of days** the pupil was absent the previous term
2) Fresh start effect: messages are sent to parents and carers **at the start of each term** and emphasise that the start of a new term is a 'fresh start'.

As this intervention is delivered to and through parents, to ensure that parents who use English as an Additional Language (EAL) can engage with the intervention, the messages will be translated into 10 languages in addition to English. The Behavioural Insights Team (BIT) developed the messages and shares these with schools in all 11 languages, giving schools the opportunity to contact parents in a non-English language. Also, schools may opt not to send messages to families, if they think the messages are not appropriate for them, which could be for a variety of reasons, such as pupil safety concerns, sensitive family situations, and so on.

Verian (formally Kantar Public UK) is conducting an impact evaluation (IE) using a two-armed cluster-RCT spanning six academic (half) terms in the intervention period across 108 schools to assess the impact of the BITUP intervention on pupil attendance rates. The evaluation also examines whether the outcomes differ by pupils' Free School Meals (FSM) and EAL status. Alongside the impact evaluation (IE), Verian is conducting a holistic implementation and process evaluation (IPE) to examine how the text messages were created and delivered; how parents receive and respond to these; and how the pupils in question react to any increase in their parents' attention to the issue of school attendance.

Design overview

The impact evaluation is designed as a two-arm randomised controlled trial of the effect of the BITUP text message intervention on Year 7 – 11 pupils' absence rates. The full description of the trial is outlined in the protocol.[2] Table 1 summarises the trial design.

Table 1: Summary of the trial design.

Trial design, including number of arms		Two-arm randomised controlled trial
Unit of randomisation		Family (siblings)
Stratification variables (if applicable)		FSM status (dichotomous) and EAL status (dichotomous)
Primary outcome	variable	Pupil absence rate
	measure (instrument, scale, source)	Attendance record in the academic year (excluding the first 5-8 week term). (Wonde) The rate is calculated as the number of sessions marked as unauthorised or authorised absence out of the total number of attendable sessions in the academic year (excluding the first 5–8 week term). In this case, a session is half a school day.
Secondary outcome(s)	variable(s)	Unauthorised pupil absence rate and authorised pupil absence rate.
	measure(s) (instrument, scale, source)	Attendance record in the academic year (excluding the first 5-8 week term). (Wonde) The unauthorised absence rate is calculated as the number of sessions marked as an unauthorised absence out of the total number of attendable sessions in period. The authorised absence rate is calculated as the number of sessions marked as an authorised absence out of the total number of attendable sessions in the period.
Baseline measure for primary outcome	measure (instrument, scale, source)	Attendance rate over the first 5–8-week term, calculated in the same manner as the primary outcome. (Wonde)

[2] BITUP_efficacytrial_protocol_v1.pdf (d2tic4wvo1iusb.cloudfront.net)

Baseline measure for secondary outcome	measure (instrument, scale, source)	Unauthorised attendance rate over the first 5–8-week term, calculated in the same manner as the first secondary outcome. (Wonde) Authorised attendance rate over the first 5–8-week term, calculated in the same manner as the second secondary outcome. (Wonde)

Research questions

Primary research question

RQ1: What is the impact of the text messaging intervention on absence rates for students aged 11–16 (Years 7–11) in treatment families who were eligible to receive the intervention compared to students aged 11–16 (Years 7–11) in control families who would have been eligible for the intervention over the whole intervention period?

Sub-group research questions

Given the focus on improving outcomes for disadvantaged pupils and on serving EAL communities, we will have the following sub-group RQs:

RQ2: What is the impact of the text messaging intervention on absence rates **for FSM-eligible (in the last 6 years)** students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the treatment group, **compared to FSM-eligible (in the last 6 years)** students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the control group?

RQ3: What is the impact of the text messaging intervention on absence rates **for EAL** students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the treatment group, **compared to EAL** students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the control group?

We also hypothesise that the impact of the intervention may vary depending on the pupil's gender or school year, as it is possible that gender and age dynamics moderate the impact of the intervention. For this reason, we include two further RQs:

RQ4: Does the impact of the text messaging intervention on absence rates **vary by gender** between students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the treatment group, compared to students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the control group?

RQ5: Does the impact of the text messaging intervention on absence rates **vary by year group** between students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the treatment group, compared to students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the control group?

The impact of the intervention could vary by initial absence rate, so we propose a sixth RQ:

RQ6: Does the impact of the text messaging intervention on absence rates **vary by level of absence in the first term of the school year** between students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the treatment group, compared to students aged 11–16 (Years 7–11) who were eligible to receive the intervention in the control group?

Secondary research questions

- What is the impact of the text messaging intervention on **unauthorised absence** rates for pupils aged 11–16 (Years 7–11) in treatment families who were eligible to receive the intervention compared to pupils aged 11–16 (Years 7–11) in control families who would have been eligible for the intervention over the whole intervention period?
- What is the impact of the text messaging intervention on **authorised absence** rates for pupils aged 11–16 (Years 7–11) in treatment families who were eligible to receive the intervention compared to pupils aged 11–16 (Years 7–11) in control families who would have been eligible for the intervention over the whole intervention period?

Randomisation

As shown in Figure 1, 980 schools were approached to participate in the trial, of which 142 schools agreed to participate. In total, 34 schools were excluded or withdrew from the trial, resulting in 108 schools being randomised (n = 87,909 families, n = 104,029 pupils).

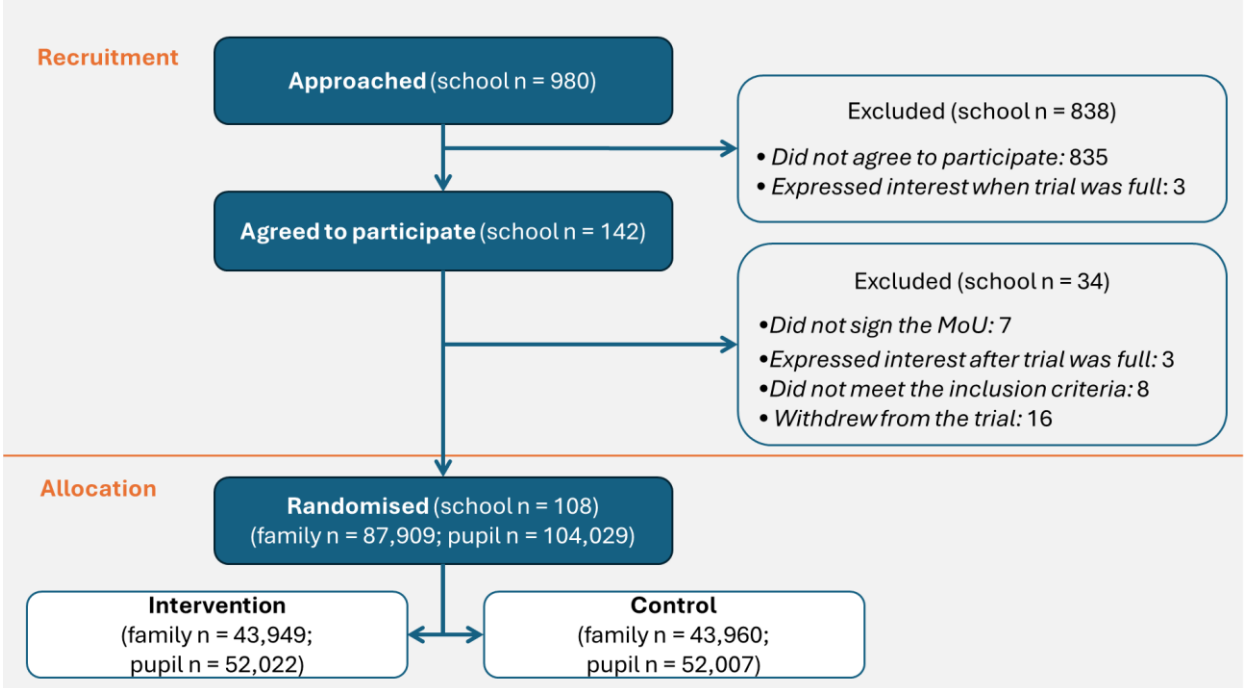

Figure 1. Trial recruitment flow diagram

Pupil data was accessed using Wonde for the first time one week before randomisation. The randomisation was done in R using package **randomizr**.[3] One researcher conducted the randomisation, then code and randomisation results were checked by another researcher. The randomisation code can be found in the 'Randomisation code' subsection of the appendix.

In total, 104,029 pupils in the 108 schools were randomly assigned into one of the two arms at the end of the first 6-week term of the 23/24 school year (19th October 2023). As set out in the trial protocol, randomisation was stratified by free school meal (FSM) eligibility and English as an additional language (EAL) status at the family level.

To mitigate spillover effects within schools and ensure that all pupils sharing a parent or guardian are in the same treatment group, pupils were randomly assigned into the treatment or control group at the family level. For this trial, families are defined as a group of siblings in the same school, as defined by

[3] https://cran.r-project.org/web/packages/randomizr/randomizr.pdf

the school. This school-level and school-defined sibling marker was used to identify 87,909 unique family IDs (an average of 1.18 pupils per family group) at randomisation.

At randomisation, 30,162 pupils (29%) were eligible for free school meals in the last 6 years (See Table 2 for a breakdown by arm). There were no missing cases of FSM status at the pupil level. For some families, 407 (0.5%) containing 871 pupils (0.8%), FSM status differed between pupils within the same family[4]. In these cases, for randomisation families were defined as FSM eligible if at least one pupil in the family was eligible for FSM. In total, 25,455 families (29%) were coded as FSM eligible.

Randomisation was also stratified by EAL status. At randomisation, for 16,809 pupils (16%) English was an additional language, for 59,319 (57%) English was not an additional language, and for the remaining 27,901 pupils (27%) the data was missing. Families were coded as EAL status if at least one sibling had EAL, otherwise families were coded as non-EAL. In total, 14,848 families (16%) were coded as EAL.

Table 2. Pupils by FSM and EAL strata.

Stratum	Families		Pupils	
	Intervention	Control	Intervention	Control
Eligible for free school meals in the last 6 years				
Yes	12,731	12,724	15,101	15,061
No	31,218	31,236	36,921	36,946
English as an additional language				
Yes	7,416	7,432	8,421	8,388
No	36,533	36,528	29,665	29,654
Missing	-	-	13,936	13,965
Total	**43,949**	**43,960**	**52,022**	**52,007**

The outcome of randomisation was communicated to the delivery team who then use this information to communicate to each school which pupils are eligible to receive a text message each term. The process is as follows:

- Before each term the delivery team use attendance data from the previous term to determine which pupils in the intervention group had an attendance rate of less than 95% and calculate the number of school days missed for these pupils.
- The delivery team summarises this information for each school in an excel spreadsheet and sends this list to them by email at the beginning of each term. This way schools are only told which pupils were in the intervention condition if they became eligible for the intervention during the trial period.

[4] The reasons for differing FSM status between siblings in this minority of observed cases are unknown and could possibly be due to admin errors.

Issues with the ways that some schools used attendance codes and how these interacted with Wonde, in ways that were not anticipated by the developer, meant that some pupils were erroneously listed as eligible (having below 95% attendance) for the intervention when they were not. Different issues occurred in waves 1 and 2. As soon as the issues were identified:

1) schools were asked to pause messaging;
2) the data problem was identified and resolved;
3) corrected lists were issued to schools;
4) messaging resumed using corrected lists.

A very small number of pupils in a very small number of schools may therefore have received the intervention in these waves when they were not eligible and the total number of missed school days may have been incorrect in a very small number of messages sent using the original list.

Primary outcome

Our primary outcome measure is the total absence rate across all terms following the preintervention period. We access this data via Wonde, a digital data-management software for schools. Wonde integrates with schools' Management Information Systems, which are used to record and manage all school data and allows schools to easily share 'live' data with partners (e.g., timetabling programmes, or EdTech apps). For this project, schools were asked to onboard with Wonde and enable data access for BIT and Verian (for the purposes of intervention delivery and evaluation respectively), meaning that for the duration of the project we can access the pre-specified and pre-agreed data variables needed for the project (which include a pupil identifier, and continuously updated per-pupil attendance data). We propose deriving the primary outcome measure using key variables captured in Wonde: the total number of absences during the intervention period and the total number of available sessions during the intervention period. This allows us to calculate a comparable rate of absence for each pupil across schools.

Secondary outcomes

Our secondary outcomes are unauthorised and authorised absence rates, as a sensitivity check of the primary outcome measure. Authorised absences are absences with permission from the school, including absences where a satisfactory explanation is provided – such as illness. Unauthorised absences are absence without permission from the school. This is important as the intervention could plausibly affect the balance of authorised and unauthorised absences as well as total absence rates.

Baseline measures

Our baseline measure is the pupils' baseline absence rate, defined as their absence rate during the pre-intervention period, specifically in the first 5-8 week term of the school year. For each outcome, the baseline absence rate is calculated using the same method as the outcome measure.

Timelines

Table 3 shows the timeline for the key milestones in the evaluation of the intervention.

Table 3: Timeline

Task	Start	End
Recruitment of settings	March 2023	September 2023
Pupil data provided by settings	July 2023	September 2023
Randomisation of Settings	19th October 2023	

Randomisation information shared with BIT	20th October 2023	
Schools send out parent messages at the beginning of each 6–8-week term	Start of the remaining 5 terms of the 23/24 academic year	
Midline IPE Qualitative Research (parents and pupils)	January 2024	March 2024
Endline IPE Qualitative Research (parents, pupils, and schools)	June 2024	July 2024
Schools & parents complete post-trial surveys	July 2024	
Verian receives final outcome data from schools (via Wonde) Wonde	July 2024	
Impact evaluation analysis & IPE analysis	August 2024	November 2024
Submission of draft EEF report	December 2024	
Final EEF report	July 2025	

Sample size calculations overview

This trial is powered to detect a Minimum Detectable Effect Size (MDES) of 0.012 standard deviations for the primary analysis of absence rate among Year 7 – 11 pupils, and 0.022 among those pupils eligible for FSM. Details of the power calculation are covered in the Updated sample size calculations section. Power calculations have been run with the package 'PowerUpR' in R.

Planned sample sizes

The trial protocol made the following assumptions when conducting power calculations to decide the sample size:

- We assumed an average of 895.17 pupils, and 518.34 families (groups of siblings attending the same school in Years 7 – 11) per school, would be recruited for the trial.[5]
- Based on a pilot, we assumed that 40.5% of pupils recruited would have an attendance rate of below 95% and therefore would be included in the impact evaluation (The Behavioural Insights Team, 2020). This implied that on average 362.54 pupils per school would be included in the impact evaluation, with 50% in the control group and 50% in the intervention group.
- We assumed that 23.8% of pupils were eligible for FSM.[6]
- We also assumed that 23.8% of pupils with attendance less than 95% would be eligible for FSM. However, we anticipated that this is likely to be a conservative estimate as pupils eligible for free FSM have higher rates of absence than other pupils at all ages (Middlemas, 2018).

The result of our calculations was that we would need 100 schools to achieve an MDES of 0.012 (see Trial Protocol).

Achieved sample size calculations

To allow for attrition we aimed to recruit 115 schools, assuming 13% would be lost from the trial after recruitment. Due to some schools withdrawing from the trial after the start of the school year, 108 schools were enrolled in the trial at the time of randomisation. BIT reported that the reasons for withdrawal were:

- Schools being affected by the issues surrounding Reinforced Autoclaved Aerated Concrete (RAAC)
- Schools deciding against sharing personal data (GDPR concerns)
- Schools deciding they no longer want to take part - most of these had decided to update all parents on days of school missed
- The members of staff that signed schools up to the project having since left the school and their replacements not wishing or not having capacity to take part

Randomisation was carried out on the 19th of October 2023 with 104,029 pupils in 87,909 families, with 43,949 families assigned to the intervention group (52,022 pupils) and 43,960 families assigned to the control group (52,007 pupils). At the time of writing this document, a further four schools have withdrawn from the trial leaving us with 105 schools. We anticipate there will be further attrition by the end of the trial period in 2024.

---

[5] There are 3,401 secondary schools in England and a total of 3,044,476 Year 7 – Year 11 pupils. This estimate comes from using National Schools Pupils and Characteristics 2021/2022: Pupil gender and year group: 2022: Browse our open data, Data catalogue – Explore education statistics – GOV.UK (explore-education-statistics.service.gov.uk).

[6] This estimate comes using National Schools Pupils and Characteristics 2022/2023: GOV.UK (explore-education-statistics.service.gov.uk). We expect pupil FSM status will be highly correlated within families. To randomise families into the treatment group by FSM status, we will treat any family as FSM eligible if at least one pupil in the family is eligible for FSM.

Updated sample size calculations

Table 4 presents the sample size calculations from the trial protocol and the updated calculations at randomisation for the full sample and the FSM subgroup. These calculations indicate the smallest effect that could be detected with 80% probability at a 5% level of significance and a set of underlying assumptions. The sample size calculations in the randomisation columns (Table 4) are based on the 108 schools retained at randomisation.

The average cluster size of 1.18 pupils per family was calculated using the actual pupil data. This is lower than the 1.73 assumed in the trial protocol. We did not have access to data on how many siblings each pupil had (i.e., single child status); we only observed siblings linked by their pupil IDs to each pupil by their school. Therefore, the sibling data we accessed was not a perfect signal of the number of siblings each pupil had at the time of randomization. For the one-pupil families, this means either there were no siblings linked to the pupil (they were only children or their siblings were not enrolled in the same school) or there was a sibling linked to the pupil but the sibling was not in one of the target year groups (Year 7 - Year 11, inclusive), or at the school's discretion, the sibling was withdrawn from the research project. Therefore, 72,887 pupils (70% of the sample) were coded as one-pupil families for the trial at randomisation, which was a higher percentage than what was assumed in the trial protocol.[7]

As in the trial protocol, we use the same assumptions for family level correlation (0.801) and school level correlation (0.0598) between baseline and endline, and school level intra-cluster correlation (0.05).[8] We also assume, as in the trial protocol, that 40.5% of pupils will have attendance rates less than 95%.

The revised sample size calculations yield an MDES of 0.012. This translates to a 0.474 percentage point change in absence rate, which is within the range of 0.3–0.5 percentage points recommended in the trial protocol. Among pupils eligible for FSM, our revised MDES of 0.022 with 108 schools is the same as our initial expectation of 0.022 with 115 schools in the trial protocol.[9] This means that we are on track to meet the estimated MDES for the evaluation, as outlined in the trial protocol.

[7] In the Trial Protocol, we made the conservative estimate that 42.49% of pupils would not have a sibling in our sample based on ONS statistics: In 2021, approximately 42.49% of families in the UK with dependent children were single child families, 42.31% were double child households, and 15.2% were three or more child households Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2021

[8] We estimated the ICC from the text message trial pilot study (BIT, 2020) and pre-post correlation using pupil-level data from a previous related evaluation (Mills et al., 2012). This gave an ICC of 0.0598 and a pre/post correlation of 0.803 for secondary schools. The time lag between pre and post-test in this case was 6 months.

[9] The actual proportion of pupils eligible for free school meals in the last 6 years in our sample (28.9%) was higher than the estimated proportion (23.8%) in the trial protocol.

Table 4: Updated sample size calculations.

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size **(MDES)**		**0.011** [0.003, 0.019]	**0.022** [0.007, 0.038]	**0.012** [0.004, 0.020]	**0.022** [0.007, 0.037]
Minimum Detectable Percentage Point Change		0.435	0.893	0.474	0.869
Pre-test/ post-correlations	Level 2 (family)	0.801	0.801	0.801	0.801
	Level 3 (school)	0.5	0.5	0.5	0.5
Intracluster correlations (ICCs)	Level 3 (school)	0.0598	0.0598	0.0598	0.0598
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
Recruited sample population					
Number of schools	Total	115	115	108	108
Average cluster size (number of pupils per family)		1.73	1.73	1.18	1.18
Families	Intervention	29,804	7,093	43,949	12,731
	Control	29,805	7,094	43,960	12,724
	Total	**59,609**	**14,187**	**87,909**	**25,455**
Pupils	Intervention	51,472	12,250	52,022	15,101
	Control	51,473	12,251	52,007	15,061
	Total	**102,945**	**24,501**	**104,029**	**30,162**
Evaluated sample population assuming 40.5% of all pupils will have attendance rate <95% in at least 1 term in the intervention period					
Average cluster size (number of pupils per family with at least 1 child <95% attendance)		1.91	1.91	1.27	1.30

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Families with at least 1 child <95% attendance	Intervention	16,850	4,010	19,688	5,762
	Control	16,850	4,011	19,689	5,745
	Total	**33,700**	**8,021**	**39,377**	**11,507**
Pupils with <95% attendance	Intervention	20,846	4,961	21,069	6,116
	Control	20,846	4,961	21,063	6,100
	Total	**41,693**	**9,923**	**42,132**	**12,216**

Analysis

Primary outcome analysis

Outcome

The primary outcome of interest is the attendance rate of a pupil across the academic year – excluding the first 5–8-week term. The first 5-8 week term was used to define those who should receive the treatment in the following term (aka if they have attendance below 95% across the prior term). The rate is calculated as the number of sessions marked as an unauthorised or an authorised absence out of the total number of available sessions to attend. This gives a consistent measure across all pupils, accounting for the fact that different schools may have different numbers of sessions in a given year. This data is collected by the Wonde platform – which links into schools' own systems – and the data recorded through Wonde is the same data that schools upload to the National Pupil Database (NPD). This means the data collected on the Wonde platform is highly accurate and aligned with school records.

Analysis

The primary analysis will be intent-to-treat using multi-level modelling to reflect the clustering of pupils (Level 1) within families (Level 2) nested within schools (Level 3). This hierarchical clustering is best modelled using a multi-level model because it will give a more precise estimate of the treatment effect compared to fixed effects with clustered adjusted standard errors, meaning we will have a higher power to detect an effect. This relies on the assumption that the predictors will be uncorrelated with the random effects. We expect this assumption to be met because the families are randomised between the control and treatment arms. Given that we have one primary research question and two groups we will not adjust for multiple hypotheses testing.

We will investigate the average impact of the intervention over the post-baseline terms by comparing the attendance of pupils in the treatment group who were eligible to receive the intervention against pupils in the control group who would have been eligible to receive the intervention, had they been in the treatment group. Pupils were eligible to receive the intervention if their attendance rate was below 95% in at least one of Terms 1–5. We propose to use the following primary model:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \varepsilon_{ijk},$$

where Y_{ijk} is the absence rate of the i^{th} pupil ($i = 1, \dots, n$, where n is the number of participating pupils) in the j^{th} family ($j = 1, \dots, m$, where m is the number of participating families) in k^{th} school ($k =$

$1, \ldots, l$, where l is the number of participating schools). $Prior_{ijk}$ is the absence rate in Term 1 for the i^{th} pupil in j^{th} family in k^{th} school, $Intervention_{jk}$ is an indicator variable for intervention allocation for the j^{th} family in k^{th} school (0 = Control, 1 = Intervention). γ_k is the deviation of school $k's$ mean z-score from the grand mean ($\gamma_k \sim N(0, \sigma_1^2)$), u_{jk} is the deviation of j^{th} family within a particular k^{th} school ($u_{jk} \sim N(0, \sigma_2^2)$), and ε_{ijk} is the residual error term for the i^{th} pupil in j^{th} family in k^{th} school ($\varepsilon_{ijk} \sim N(0, \sigma_3^2)$). The multi-level nature of this equation (u_{jk}) accounts for the randomisation within school. Note that in this model and in the models that follow we do not expect to transform or scale variables. In addition, in the case that our proposed models do not converge, we will simplify the models by removing family-level nesting. In which case, the models will nest pupils within schools.

We include controls for our pupil level stratification variables, FSM (using FSM_EVER_6, coded in Wonde as *extended_details.data.free_school_meals_6*) and EAL (using the variable coded in Wonde as *extended_details.data.english_as_additional_language*). These controls will be either binary stratification variables at the pupil level or categorical variables with three levels (for example, FSM, non-FSM, or not known) in the case of missing data for a pupil.

Secondary outcome analysis

Outcome

The secondary outcomes of interest are the (1) authorised and (2) unauthorised absence rates. Like in the primary outcome analysis, we will calculate the number of sessions marked as (1) authorised and (2) unauthorised absence out of the total number of attendable sessions each term. Again, this data will be collected via the Wonde data systems.

Authorised absences are absences with permission from the school, including absences where a satisfactory explanation is provided – such as illness. Unauthorised absences are absence without a permission from the school. Inclusion of both is important as the intervention could plausibly affect the balance of authorised and unauthorised absences as well as total absence rates.

Analysis

The specification of these models is the same as for the model of the primary outcome – with the coefficient for the absence rate in the first term referring to either the authorised or unauthorised absence rate, respectively. As before, this will be intent-to-treat analysis using multi-level modelling to reflect the clustering of pupils within families nested within schools.

The model for authorised absences is:

$$Y^A{}_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 \, Prior^A_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \varepsilon_{ijk},$$

where Y^A_{ijk} is the authorised absence rate of the i^{th} pupil ($i = 1, \ldots, n$, where n is the number of participating pupils) in the j^{th} family ($j = 1, \ldots, m$, where m is the number of participating families) in k^{th} school ($k = 1, \ldots, l$, where l is the number of participating schools). $Prior^A_{ijk}$ is the authorised absence rate in Term 1 for the i^{th} pupil in j^{th} family in k^{th} school, $Intervention_{jk}$ is an indicator variable for intervention allocation for the j^{th} family in k^{th} school (0 = Control, 1 = Intervention).

The model for unauthorised absences:

$$Y^{UA}{}_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 \, Prior^{UA}_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \varepsilon_{ijk},$$

where Y^{UA}_{ijk} is the unauthorised absence rate of the i^{th} pupil ($i = 1, \ldots, n$, where n is the number of participating pupils) in the j^{th} family ($j = 1, \ldots, m$, where m is the number of participating families) in k^{th} school ($k = 1, \ldots, l$, where l is the number of participating schools). $Prior^{UA}_{ijk}$ is the unauthorised

absence rate in Term 1 for the i^{th} pupil in j^{th} family in k^{th} school, $Intervention_{jk}$ is an indicator variable for intervention allocation for the j^{th} family in k^{th} school (0 = Control, 1 = Intervention).

For both γ_k is the deviation of school $k's$ mean z-score from the grand mean ($\gamma_k \sim N(0, \sigma_1^2)$), u_{jk} is the deviation of j^{th} family within a particular k^{th} school ($u_{jk} \sim N(0, \sigma_2^2)$), and ε_{ijk} is the residual error term for the i^{th} pupil in j^{th} family in k^{th} school ($\varepsilon_{ijk} \sim N(0, \sigma_3^2)$). The multi-level nature of this equation (u_{jk}) accounts for the randomisation within school. In addition, we will include controls for our pupil level stratification variables, FSM and EAL as in the primary outcome analysis.

Subgroup analyses

We will examine the impact of the intervention on our primary outcome for five subgroups of interest. These are FSM (RQ2), EAL (RQ3), Gender (RQ4), Year Group (RQ5), and absence level in the first term of the school year **(**RQ6**)**. The first three of these are binary indicators, with FSM=1 indicating the pupil is eligible for FSM (or has been eligible in the last 6 years), EAL=1 indicating the pupil is identified by the school as speaking English as an additional language, and gender indicating if the pupil is male or female. For year group, we have five-year groups taking part in the intervention – Year 7, Year 8, Year 9, Year 10 and Year 11. Initial absence level in the first term will be treated as a continuous measure.

For each dichotomous subgroup (FSM, EAL, and Gender) we will run a model with an interaction term:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \beta_5 \left(Intervention_{jk} . X_{ijk}\right) + \varepsilon_{ijk},$$

where X_{ijk} is equal to *FSM*, *EAL* or *Gender* at the pupil level, respectively. (Note that in the case of Gender, we will also add Gender fixed effect to the model.) This model will allow us to examine the effects of the intervention by each subgroup, as well as determine whether there are significant differences between the subgroups.

In line with EEF Analysis guidance (Education Endowment Foundation, 2022), we will also produce split sample analysis results in the appendix of the evaluation report for each subgroup, containing only pupils from that subgroup.

To address the RQ5, the model will be:

$$\begin{aligned} Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} \\ + \beta_5 \left(Intervention_{jk} . Yr8_{ijk}\right) + \beta_6 \left(Intervention_{jk} . Yr9_{ijk}\right) \\ + \beta_7 \left(Intervention_{jk} . Yr10_{ijk}\right) + \beta_8 \left(Intervention_{jk} . Yr11_{ijk}\right) + \varepsilon_{ijk}. \end{aligned}$$

In this model, Year 7 is designated as the reference category. This means that the coefficients for the interaction between the intervention and each year group (Year 8 to Year 11) are interpreted in relation to the baseline established by Year 7. The reference category, Year 7, will not be explicitly included in the model as its effect are now accounted for in the intercept (β_0). We will also produce split sample analysis results in the appendix of the evaluation report for each year group, containing only pupils from that year group.

To address RQ6, the model will be:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \beta_5 (Intervention_{jk} . Prior_{ijk}) + \varepsilon_{ijk}.$$

Additional analyses

First time of receiving a text

The capacity for an intervention to maintain its effects over time is a critical component of behavioural intervention. However, across contexts little is known about whether the effects of behavioural interventions persist over time, and there is some evidence that over time effects may diminish (Hagger & Weed, 2019; Hecht et al., 2019; Willcox et al., 2019). One hypothesis is that, across the trial period, the intervention will have a larger effect on attendance the first time that it is received (i.e., a novelty effect).

We expect that the more effective the intervention is the less likely a treated pupil should be to become eligible for the intervention in subsequent terms. Nevertheless, we expect some individuals will receive multiple texts during the trial. Therefore, we will conduct an additional analysis to investigate the impact of receiving the treatment for the first time on absence rates in the subsequent term.

This is answering the research question:

What is the impact of the text messaging intervention on absence rates for pupils aged 11–16 (Years 7–11) in treatment families in the first term they were eligible to receive the intervention compared to pupils aged 11–16 (Years 7–11) in control families in the first term they would have been eligible for the intervention?

The proposed model is:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \varepsilon_{ijk},$$

where Y_{ijk} is the absence rate of the i^{th} pupil ($i = 1, \dots, n$, where n is the number of participating pupils) in the j^{th} family ($j = 1, \dots, m$, where m is the number of participating families) in k^{th} school ($k = 1, \dots, l$, where l is the number of participating schools), in the first term a pupil's family is (or would be, in the case of control) eligible to receive the intervention. The rest of the terms are the same as in the primary analysis model.

Depending on the term

There is evidence that absenteeism can vary with time of year, with a winter peak in absenteeism, possibly due to higher incidence of sickness (Zerbini et al., 2019). As a robustness check to examine seasonal trends we will extend our primary model to examine how the intervention effect changes over intervention terms. This answers the following question:

- Does the impact of the text messaging intervention on absence rates vary between the school terms (5-8 week periods) for secondary school pupils aged 11–16 (Years 7–11) in treatment families eligible to receive the intervention compared to secondary school pupils aged 11–16 (Years 7–11) in control families who would have been eligible for the intervention?

The extended model is:

$$Y_{tijk} = \beta_0 + \gamma_k + u_{jk} + \tau_{ijk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jkt} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \beta_5 Term_t + \beta_6 Intervention * Term_{tjk} + \varepsilon_{tijk},$$

where $Term_t$ is categorical variable indicating t^{th} post-intervention term ($t = 1, \dots, 5$ post-intervention terms), $Intervention * Term_{tjk}$ is an interaction term indicating allocation for the j^{th} family in k^{th} school in t^{th} post-intervention term (0 = Control, 1 = Intervention), and τ_{ijk} is a random effect for each pupil, capturing the individual-specific variation that is consistent across different terms,

School level factors

As schools vary across observable and unobservable factors (including ones that may impact attendance levels, such as the approaches they take to improving attendance) it is reasonable to assume that eligibility for the intervention and the effectiveness of the intervention may vary across schools. As a robustness check to control for this, we will add a school level variable for the proportion of pupils of that school who became eligible to receive a text message:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \beta_5 SPPE_k + \varepsilon_{ijk},$$

where $SPPE_k$ specified the proportion of pupils in k^{th} school who became eligible to receive a text message.

Dosage

Our ability to conduct a meaningful dosage evaluation for this intervention is severely limited as the effectiveness of the intervention interacts with eligibility during multiple terms (dosage). If the intervention is effective, this would reduce the number of messages families get on average. If families receive more than one dose of the treatment this could mean the intervention did not have an impact, but it could also be the case that the intervention had an effect but not one that was great enough to raise the attendance of children from that family to at least 95%. Therefore, examining the effects of dosage means specifically looking for the effects of dosage only among the subsample of treated families for whom receiving the intervention the first time did not increase the attendance rate to at least 95% in all subsequent terms.

As dosage is a key potential feature of this intervention (parents may receive multiple messages across the year) we will conduct a dosage analysis, albeit with significant caveats. The families included in our dosage analysis (looking at >1 dose of intervention) will be those whose child's attendance rate did not increase to at least 95% in subsequent terms following the intervention. If the intervention has a positive impact we expect the group receiving multiple doses of the messaging to be smaller than that of families receiving one dose. This discrepancy will be greater the more effective the intervention is.

Based on the reasons outlined above, we anticipate a higher level of uncertainty associated with the intervention's dosage as it increases because fewer pupils are eligible for the intervention. Despite this, we include a dosage analysis aimed at answering the following research question:

- Does the impact of the text messaging intervention on absence rates for secondary school pupils aged 11–16 (Years 7–11) in treatment families who received the intervention compared to secondary school pupils aged 11–16 (Years 7–11) in control families who would have been eligible for the intervention vary by dosage?

The extended proposed model is:

$$Y_{ijk} = \beta_0 + \gamma_k + u_{jk} + \beta_1 Prior_{ijk} + \beta_2 Intervention_{jk} + \beta_3 FSM_{ijk} + \beta_4 EAL_{ijk} + \beta_5 Dose_{ijk} + \beta_6 (Intervention_{jk} . Dose_{ijk}) + \varepsilon_{ijk},$$

where $Dose_{ijk}$ is a categorical variable indicating the number of doses received by pupils who received more than 1 dose (e.g., 2 doses, 3 doses, 4 doses, etc.).

Additional outcomes

We will also try to understand the extent to which the intervention impacted on the accuracy of parents' perception of their child's absence. The parent survey, which will be sent to parents of pupils who are or would have been eligible to receive the intervention (control and treatment), will ask parents to estimate their child's absence in number of days over the past two terms (as we felt recollection was less likely to be accurate if we ask them to estimate for a longer period). This will provide an indicative answer to the question:

- What is the impact of the text messaging intervention on the accuracy of parental knowledge about their child's absence?

We will explore this descriptively in the analysis, by reporting summary statistics for this variable in treatment and control. This is because we do not have a baseline survey in our design (as it would have been perceived to have been almost a stronger prime for parents to pay attention to attendance), the return rates are likely to be low and biased (e.g., the sample is unlikely to be representative due to non-random self-selection).

Longitudinal follow-up analyses

There is no longitudinal follow-up as part of this trial, but data will be archived to facilitate future longitudinal analyses.

Imbalance at baseline

A well-conducted randomisation ideally should create groups that are equivalent at baseline (at the point of randomisation), with any imbalance at baseline occurring by chance. However, to check for, and monitor, imbalance at baseline in the realised randomisation, baseline equivalence testing will be conducted at the pupil level.

At the family level, we will check the balance in the following variables by means of cross-tabulations and histograms that assess the distribution of each characteristic between the control and intervention groups:

- Proportion of families eligible for FSM
- Proportion of families classified as EAL
- Number of siblings (in the year 7-11)

Note that we will not undertake statistical testing as the initial randomisation necessarily means that any imbalance we observe is due to chance and therefore statistical tests are not appropriate.

At the pupil level, balance will be assessed as above for the following characteristics:

- FSM
- EAL
- Gender
- SEND
- Ethnicity
- Age
- Year group

In addition, in line with EEF's efficacy reporting template (Education Endowment Fund, 2019), we will report effect sizes to quantify the magnitude of any observed imbalance in pupils' baseline attendance rate between the intervention and control groups.

School level is not important for balance because we are randomising within schools, however we will report characteristics of our schools in our sample from the latest available year. This will include:

- Geographic region
- School size (number of pupils)
- Proportion FSM (whole school)
- Proportion EAL (whole school)
- Absence rate (average annual)

Missing data

Attrition across both trial arms will be explored as a basic step to assess bias across arms. We will describe the extent of missingness of all variables to be used as part of the analysis; both overall and then by arm. We will provide cross-tabulations of the proportions of missing values on all characteristics used in the analysis (at both pupil, family and school level), as well as on the primary outcome measure and secondary outcome measures. It is most likely in this case that missingness will be related to a school level issue and we can show the average level of missingness per school and map out any outliers.

For less than 5% missingness overall in our primary/secondary outcome we propose a complete-case analysis. In cases where there is over 5% of missingness we will explore the missingness mechanisms as detailed below.

Missing information generally is classified into the following situations:

- 'Missing completely at random' (MCAR): the likelihood of data being missing(/present) for an indicator is completely due to chance, and does not depend on other factors or on the extent/magnitude of the thing that is being measured.
- 'Missing at random' (MAR): the likelihood of data being missing for an indicator does depend on other factors but does not depend on the extent/magnitude of the thing that is being measured.
- 'Not missing at random' (NMAR); the likelihood of data being missing for an indicator depends on the extent/magnitude of the thing that is being measured.

These different situations determine which particular analytical approaches may be suitable. Undertaking analysis on pupils with information for every single indicator and omitting any other pupils from the analysis assumes that missing data are 'Missing completely at random' (MCAR). Making inferences/estimates of any missing information assumes that missing data are 'Missing completely at random' (MCAR) or 'Missing at random' (MAR). It is generally highly unlikely that in the case of a large amount of missing data (over 5%) that it is MCAR but if we consider this to be the case based on our inferences around the patterns of missingness we will only complete a complete case analysis.

To consider if data could be MCAR or MAR there are a number of tests that will be conducted. Here, by first examining the summary statistics and cross-tabulations, as well as any other information we have about why data is missing, it may be possible to either assume or rule out MCAR.

Second, we will estimate a model predicting missingness to examine whether the covariates in our analysis model jointly (using an F-test) predict the absence of our outcome data. In this scenario we model missingness (defining individuals with missing primary outcome data at endline) as a function of baseline covariates (e.g., size of a school and the proportion of FSM eligible pupils at the school level). The analysis model for this approach will mirror the multilevel level model specified in the primary analysis section with pupils clustered in families in schools, but the outcome will be a binary variable identifying missingness (where 1=missing; 0=complete) in a multilevel mixed-effects logistic regression model. We can similarly model missingness at baseline (defined as pupils with missing baseline data, namely first term absences) as a function of endline covariates, including treatment. If any differences are found, significant at $p < 0.05$, we will supplement the analysis using multilevel multiple imputation using package 'mice' in R.[10] Results will be reported alongside the complete case analysis.

[10] CRAN - Package mice (r-project.org)

Compliance

We will define compliance at the pupil level. In practice, compliance is determined by the schools' actions of sending texts to the correct parents in the correct time frame.[11] Given we have randomised at family level, a situation could arise where two siblings with attendance less than 95% are eligible to be treated, but the parent only receives the text about one sibling in the correct time frame (compliant) and in error, doesn't receive the text about the other sibling (or not in the correct time frame) and thus are not compliant. In this scenario, the parent has received prompts about one child, but not the other.

This is further complicated by the fact that our indicator of compliance (receiving the text message) is only valid in the first term in which the text message is received (the second term of the 23/24 academic year). In subsequent terms, this indicator becomes 'treatment inherent', meaning that the likelihood of receiving a text message is influenced by the effectiveness of the treatment itself. If the text message successfully changes behaviour in some cases it could improve a pupil's attendance to above 95%, meaning the pupil will no longer be eligible to receive further texts in subsequent terms. Therefore, in these later terms, receiving a text message is not just a matter of compliance but in some cases could also be an outcome influenced by the treatment's success.

Ideally, pupils would be considered compliant if the school sends the correct text message to the correct family in the correct time frame. Given the intervention is considered to have two key components – the number of days missed and sent within the first two weeks of term (to harness the fresh start effect) – we would ideally want our measure of compliance to include both components. However, without auditing all the messages that schools send, we cannot know if the correct text message is sent out (i.e., if the number of days is correct).

We will ask schools to tell us the date on which they sent the text message to the families of each of the pupils marked as eligible to receive a text message for that term, which will enable us to infer if a text is sent within the correct time frame. However, this data may be inaccurate (e.g., schools may input the same date into all cells to save time even if not all texts were sent on the same date) or indeed incomplete (this ask may place too high a burden on schools). As such, we propose to use two binary compliance indicators.

Primary indicator (receiving the text):

- Compliance = 0 if a pupil's family does not receive a text message when the pupil has been considered eligible to do so during the term in which they were eligible
- Compliance = 1 if a pupil's family does receive a text message when the pupil has been considered eligible.

Secondary indicator (timeliness):

- Compliance = 0 if the text message is sent more than two weeks after the start of term
- Compliance = 1 if the text message is sent within the first two weeks of term.

While the threshold of two weeks is arbitrary, the Delivery Team (BIT) believe this is a suitable threshold for the hypothesised fresh start to occur. If schools do not provide us with the date texts are sent or if this data is considered incomplete/inaccurate, we will only use the primary compliance indicator. Conditional on receiving monitoring forms from schools, we would rather not exclude individuals who we have data on simply due to missing dates and thus reduce our sample. To effectively use the secondary indicator of compliance, we have set a high standard for the completeness of date data. Specifically, we aim to construct this measure for at least 90% of the

[11] The published Trial Protocol containing further information can be found here.

sample for which we can create a binary measure. This approach ensures that we do not unnecessarily exclude individuals from our analysis due to missing date information, thereby maintaining a more robust sample.

We are also aware there is a risk that we will not be able to construct the binary measures for our sample of interest – either through schools not completing and returning the monitoring forms, or large amounts of missing data in the monitoring forms. This may result in a very restricted sample that we can run the analysis on, which will make comparisons to our main ITT estimates difficult and may result in selection biases in that certain types of schools may be better at sending texts and better at completing monitoring forms. We do not think it appropriate to assume that people have not been treated (text message compliance indicator=0) if we do not receive a monitoring form. Thus, we will run the analysis on the sample we have but will report on the sources of missingness in constructing our compliance measures and interpret any such findings with appropriate caveats.

We have begun to receive monitoring data from schools. For ~50% of our sample, schools have returned the forms, indicating the specific dates when messages were sent out. This column was pre-populated by BIT in the forms sent to schools to reduce school burden, although we note that this could impact on the accuracy of the data provided by schools (as they may simply not have changed it). Thus, from the fifth intervention wave we asked BIT not to pre-populate this column, so that we can have more confidence in the date inputted by schools. Our measures of compliance also capture non-compliance in the form of schools forgetting to send text messages, mistakes in sending out the messages, and discretionary removal of pupils from the eligible list of pupils.

We will include non-compliance in the other direction – where eligible control pupils receive the text message in error – if schools provide us with the data on who they texted and that includes pupils not on the list of eligible pupils. The non-compliance in the other direction will take a similar form to that above, following BIT's recommendation from their formative research with schools. Schools will receive a list of eligible pupils and be asked to text them and record the date they send the text message. They will then be asked to share with Verian a list of all the pupils whose parents were sent a message, and the date on which this was sent. If this list includes additional pupils, we will use this data to complete a two-way compliance analysis as our preferred approach.

Some amount of caution is required, given that some schools may simply not tell us if they text pupils who are not on the list of those eligible to receive a message. Based on the formative work conducted with schools, BIT has suggested that the reporting to Verian will happen in the same CSV document that BIT shares with schools, which will contain the list of families of pupils that are eligible to receive a text each term. While this approach eases the reporting workload for schools, it also heightens the likelihood of inaccuracies in reporting compliance data. Specifically, there is an increased risk that schools might not report instances where families who are not on the eligible list were mistakenly sent text messages. We note the risk of data inaccuracy and/or incompleteness. Our compliance measures here will follow a similar pattern to those listed above but for eligible control pupils as opposed to treated pupils.

We will use an instrumental variable approach using Two Stage Least Squares (2SLS) to estimate the Compliance Average Causal Effect (CACE) using the primary outcome only. This analysis will also only be conducted using absence rates during the term of data after the first text message is received (at an individual level) as the outcome.

As mentioned previously, the primary measure of compliance that assesses whether a family received a text is invalid in any terms following the term in which a family first received an intervention text as compliance has then become treatment inherent. To ensure the CACE estimator is comparable to the ITT estimator, we will compare our result to the ITT analysis using the same approach; we will generate and compare the two estimators based on data only from the first term in which they receive their first text message. This ITT estimation will be run as a robustness check. We will not compare

the CACE estimator to the primary research question estimate using the full year absence rate as an outcome due to the treatment inherent nature of this.

Intra-cluster correlations (ICCs)

We will report the school-level and family-level ICCs based on the primary outcome measure. These will be calculated using the primary outcome analysis model, and the primary outcome analysis model with no predictors, accounting for the clustering of pupils in families and families within schools (the so-called empty model).

Effect size calculation

We will estimate the effect size of the intervention using an adaptation of Hedges' g (Hedges, 2007), as done in previous EEF efficacy trials involving measures of reading attainment, comprehension, and fluency (Dimova & Illie, 2021). Specifically,

$$g = \frac{(\hat{Y}_T - \hat{Y}_C)_{adjusted}}{\sqrt{\sigma_s^2 + \sigma_f^2 + \sigma_{error}^2}}$$

where $(\hat{Y}_T - \hat{Y}_C)_{adjusted}$ is the mean difference between treatment and control groups adjusted for baseline attendance rate and $\sqrt{\sigma_s^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (school level σ_s^2, family level σ_f^2, and individual level σ_{error}^2 variance).

We will calculate confidence intervals for the effect size using

$$g - c_{\alpha/2}\gamma_g \leq \hat{g} \leq g + c_{\alpha/2}\gamma_g$$

where $\hat{g}$ is the estimated effect size, and γ_g is the estimated variance, and $c_{\alpha/2}$ is the 100(1- $\alpha/2$) percentage point of the standard normal distribution.

Conditional addition to the analysis approach

The primary outcome is the attendance rate over the entire academic year. However, this measure may dilute the intervention's effect, as it includes the pre-treatment period in the post-treatment outcome. For instance, if a pupil receives the text message in Term 5, the average attendance rate includes Term 5 and all prior terms.

If the primary analysis for RQ1 shows no evidence of the intervention's effectiveness, but the additional analysis on the first receipt of the text does, we will conduct an additional exploratory analysis that examines the average effect of the intervention from the term of the first text received onwards, focusing on terms after the child's attendance first falls below 95%. This will only be done for the total sample (RQ1). This additional analysis would provide exploratory evidence that could help us interpret differences between the overall impact of the intervention and the seeming impact in the first term after receiving a message.

References

Dimova, S., & Illie, S. (2021). *Peer Assisted Learning Strategies UK: Statistical Analysis Plan.* PALS_SAP_update_2021.pdf (d2tic4wvo1iusb.cloudfront.net)

Education Endowment Foundation. (2022, October). *Statistical analysis guidance for EEF evaluations*. https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1677842964

Education Endowment Fund. (2019). *EEF Evaluation Report Template*. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fd2tic4wvo1iusb.cloudfront.net%2Fproduction%2Fdocuments%2Fevaluation%2Freporting-templates%2FEEF_evaluation_report_template_2019.docx%3Fv%3D1708344032&wdOrigin=BROWSELINK

Hagger, M. S., & Weed, M. (2019). DEBATE: Do interventions based on behavioral theory work in the real world? *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 36. https://doi.org/10.1186/s12966-019-0795-4

Hecht, C. A., Priniski, S. J., & Harackiewicz, J. M. (2019). Understanding Long-Term Effects of Motivation Interventions in a Changing World. *Advances in Motivation and Achievement : A Research Annual*, *20*, 81–98. https://doi.org/10.1108/S0749-742320190000020005

Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavourial Sciences*, *34*(4), 341–370. https://doi.org/10.3102/1076998606298043

Middlemas, J. (2018). Absence rates by gender, age and free school meal status. *Department for Education*.

The Behavioural Insights Team. (2020, July 17). *Improving pupil attendance through timely nudges*. https://www.bi.team/blogs/improving-pupil-attendance-through-timely-nudges/

Willcox, J. C., Dobson, R., & Whittaker, R. (2019). Old-Fashioned Technology in the Era of “Bling”: Is There a Future for Text Messaging in Health Care? *Journal of Medical Internet Research*, *21*(12), e16630. https://doi.org/10.2196/16630

Zerbini, G., van der Vinne, V., Otto, L. K. M., Monecke, S., Kantermann, T., & Merrow, M. (2019). Tardiness Increases in Winter: Evidence for Annual Rhythms in Humans. *Journal of Biological Rhythms*, *34*(6), 672–679. https://doi.org/10.1177/0748730419876781

Appendix

Randomisation code

```{r, label="setup"}

rm(list = ls())


library(tidyverse)

library(readxl)

library(randomizr)

library(table1)

library(data.table)

library(reshape2)


setwd("//kt.group.local/root/EMEA-UK-SLOUGH-1643/kantar/KI/Social_Research/1.CLIENTS/EEF/262400452 BITUP - Updating Parents by Text/12. IE data (SERVER ONLY - RESTRICTED)/Randomisation")


```

There are 108 schools that have approved access to their data.

```{r, label="Load school list"}

#get the list of school names used when pulling the student data via the Wonde API

school_list<-read_excel("All-Audited-Schools.xlsx")


#get the list of schools that gave approval to access data

school_approval<-read_excel("[Shared with KP] - BITUP sample - Wonde actions confirmed.xlsx")

school_confirm <- subset(school_approval, school_approval$`Confirmed Wonde withdrawals complete?`=="Yes")

nrow(school_confirm)

sum(is.na(school_confirm$URN))


school_confirm <- merge(school_list, school_confirm, by.x="region.identifiers.urn", by.y="URN", all.y=TRUE)

nrow(school_confirm)
```

#get the list of school names - 108 schools

all_school <- school_confirm$name

all_school

```

Read the students data for all the 108 schools.

```{r, label="Load student data (pooling all schools)"}

#read student data from all schools and store in data

data <- data.frame(matrix(nrow=0,ncol=0))

datapath <- "//kt.group.local/root/EMEA-UK-SLOUGH-1643/kantar/KI/Social_Research/1.CLIENTS/EEF/262400452 BITUP - Updating Parents by Text/12. IE data (SERVER ONLY - RESTRICTED)/Trial data/"

for (school in all_school) {

  if (file.exists(paste0(datapath,school,"-Students.xlsx"))==TRUE){

  temp <- read_excel(paste0(datapath,school,"-Students.xlsx"), col_types = "text")

  temp$school <- school

  data <- bind_rows(data, temp)

  }

}

#check number of schools in the combined data

length(unique(data$school))

#Check variables in combined data

sjPlot::view_df(data, file = "student_var_list.html", show.type = TRUE, show.na = TRUE, show.prc = TRUE)

#Create a copy of the raw data

data0 <- data
```
```

```

Check if there are multiple entries for the same student - No duplicate rows or multiple rows with same student id.

There are 116420 students in the raw data from the 108 schools.

```{r, label="Multiple entries for the same student"}

#remove two rows that are completely the same

nrow(data)

data <- data %>% distinct()

nrow(data) #check if it's different from the number above

#list entries for the same id that are not completely the same - students in different schools cannot have the same id

student_duplicates <- data %>% group_by(id) %>% filter(n() > 1)

nrow(student_duplicates)

```

Based on the following note from the API manuals <https://docs.wonde.com/docs/api/sync/#student-object>: Students will have upi if they are enrolled - as no one has missing upi in the data, we will assume they are all enrolled currently

"Unique Person Identifier - This field is the mis_id and school_id combined to create a unique hash. There are benefits of using the UPI when matching users, for example, when a student is disenrolled the student will be removed from Wonde. If that student is then re-enrolled the Wonde ID will change but the UPI will remain the same."

```{r, label="Check UPI"}

sum(is.na(data$upi))

```

Check if the range of number of students per school in raw data looks sensible - school sizes between 221 and 2515
```

```{r, label="Number of students per school"}

school_size <- as.data.frame(table(data$school,useNA="always"))

school_size[is.na(school_size$Var1)==TRUE,]$Freq

summary(school_size[is.na(school_size$Var1)==FALSE,]$Freq)

```

Excluding students not in year 7-11 - There are 44 students without a value for school year. They are spread out in 10+school. Students in the data who do not have a value for school year being 7-11 are excluded. This leads to exclusion of 116420- 104029= 12391 students.

```{r, label="Restricting to year 7-11"}

table(data$education_details.data.current_nc_year,useNA="always")

table(subset(data,is.na(data$education_details.data.current_nc_year)==TRUE)$school)

data<-subset(data, data$education_details.data.current_nc_year == "7"
        | data$education_details.data.current_nc_year == "8"
        | data$education_details.data.current_nc_year == "9"
        | data$education_details.data.current_nc_year == "10"
        | data$education_details.data.current_nc_year == "11")

length(unique(data$school))

nrow(data)

data1<-data #create a copy of the screened data

```

There are students who are not in year 7-11 in 60 of the 108 schools. The number of students not in year 7-11 in a school varies between 2 to 720.

```{r, label="Number of students in years other than 7-11 per school"}
```

```
data_not_year7to11 <-subset(data0, data0$education_details.data.current_nc_year != "7"
        & data0$education_details.data.current_nc_year != "8"
        & data0$education_details.data.current_nc_year != "9"
        & data0$education_details.data.current_nc_year != "10"
        & data0$education_details.data.current_nc_year != "11")

data_not_year7to11 <- as.data.frame(table(data_not_year7to11$school))
length(unique(data_not_year7to11$Var1))
summary(data_not_year7to11$Freq)
```

Check number of students per school after exclusion - school size varies between 221 to 1795 after exclusion.

```{r, label="Number of students per school after exclusion"}

school_size <- as.data.frame(table(data$school,useNA="always"))
school_size[is.na(school_size$Var1)==TRUE,]$Freq
summary(school_size[is.na(school_size$Var1)==FALSE,]$Freq)
tail(sort(school_size$Freq),10)
```

Check number of missing values for key variables per school:

- no missing values for student id, sibling data, and FSM
- percentage of missing values for EAL in a school varies from 0% to 94%

```{r, label="Missing values for key variables"}
sum(is.na(data$id))
sum(is.na(data$original.siblings.data))
```

```
missing_values <- data %>% group_by(school) %>% summarize(
  missing_share_EAL = mean(is.na(extended_details.data.english_as_additional_language)),
  missing_share_FSM = mean(is.na(extended_details.data.free_school_meals_6)),
  missing_count_EAL = sum(is.na(extended_details.data.english_as_additional_language)),
  missing_count_FSM = sum(is.na(extended_details.data.free_school_meals_6)),
  school_size = length(unique(id)))

write.csv(missing_values, "missing_values_year7-11.csv")

summary(missing_values$missing_share_EAL)
summary(missing_values$missing_share_FSM)
```

Percentage of students with FSM status in a school ranges from 6.8% to 65.3%.

Percentage of students with EAL status (excluding missing values) in a school ranges from 1.6% to 91.7%.

```{r, label="Check and clean stratification variables"}
#recoding EAL and FSM from T/F to 1/0
data$EAL <- if_else(data$extended_details.data.english_as_additional_language=="FALSE", 0, 1)
data$FSM <- if_else(data$extended_details.data.free_school_meals_6=="FALSE", 0, 1)

table(data$EAL,useNA="always")
table(data$FSM,useNA="always")

#check if percentage of EAL, FSM looks right for each school
pc_EAL_FSM <- aggregate(cbind(EAL,FSM)~school, data=data, mean,na.action = na.omit)

summary(pc_EAL_FSM$EAL)
```

summary(pc_EAL_FSM$FSM)

```

In the raw data for those in year 7-11, 67793 students have no siblings, 31197 have one sibling, while one student has seven siblings (maximum). Mean number of siblings for each student in a school ranges from 0.037 to 0.778.

```{r, label="Getting the list of siblings for those who have any siblings"}

#getting the list of sibling ids from the original.siblings.data variable

pattern <- "student': '(.*?)', 'data_of_birth"

data$sibling_all_id <- gregexpr(pattern, data$original.siblings.data, perl = TRUE)

data$sibling_all_id <- regmatches(data$original.siblings.data, data$sibling_all_id)

clean_text <- function(text) {

  gsub(pattern, "\\1", text)

}

data$sibling_all_id <- lapply(data$sibling_all_id, clean_text)

#creating the number of siblings variable

data$sibling_num <-  lapply(data$sibling_all_id, length)

data$sibling_num <- as.numeric(data$sibling_num)

#distribution of number of siblings in raw data

table(data$sibling_num, useNA="always")

#distribution of number of siblings per school

sibling_school <- data %>% group_by(school) %>%

  summarize(mean = mean(sibling_num),

            min = min(sibling_num),

            quartile1 = quantile(sibling_num, probs = c(0.25)),

            median = median(sibling_num),

```
        quartile3 = quantile(sibling_num, probs = c(0.75)),
        max = max(sibling_num))

summary(sibling_school$mean)
```

There are 41873 siblings listed in the student data, 554 of which are duplicates, leaving 41319 unique cases.

There are 37293 unique sibling ids among the 41319 cases. Out of them, 6170 unique sibling ids (6978 rows) do not exist in the student data.

For those siblings who exist in the student data, none of them belong to a different school.

```{r, label="Checking if siblings are in the same school"}
#long data of student and each of their sibling
temp <- subset(data,select = c("id","school","EAL", "FSM", "sibling_all_id"))
sibling <- unnest(data, sibling_all_id)
nrow(sibling)

#remove duplicate entries
sibling <- sibling %>% distinct()
nrow(sibling)
length(unique(sibling$sibling_all_id))

#get student data of sibling
temp <- subset(data,select = c("id","school","EAL", "FSM"))
colnames(temp)[colnames(temp) == "school"] <- "school_sibl"
colnames(temp)[colnames(temp) == "EAL"] <- "EAL_sibl"
colnames(temp)[colnames(temp) == "FSM"] <- "FSM_sibl"
sibling <- merge(sibling, temp, by.x="sibling_all_id", by.y="id", all.x=TRUE)
```

```
#check if siblings do not exist in the student data

temp <- subset(sibling, select=c("id","school","sibling_all_id","school_sibl"))

sibling_school_NA <- subset(temp, is.na(temp$school_sibl))

nrow(sibling_school_NA)

length(unique(sibling_school_NA$sibling_all_id))


#check if siblings are in the same school

sibling_school_diff <- subset(temp, is.na(temp$school_sibl)==FALSE)

sibling_school_diff <- subset(sibling_school_diff,
!(sibling_school_diff$school==sibling_school_diff$school_sibl))

nrow(sibling_school_diff)
```

All listed siblings are marked as on-roll in the sibling variables.

```{r, label="Checking the on-roll status of the siblings that do not exist in student data"}

#getting the variables for individual siblings in student data

data0_sibling <- data1 %>%

   select(id, starts_with('siblings_')) %>%

   select(id, ends_with('_student') | ends_with('_on_roll') )


#reshape to long form

data0_sibling <-
data.table::melt(setDT(data0_sibling),id="id",variable.name="sibling",measure=patterns(

                sibling_id=".*_student",

                sibling_on_roll=".*_on_roll"))


#remove NA rows

data0_sibling <- subset(data0_sibling, is.na(data0_sibling$sibling_id)==FALSE)

str(data0_sibling)

length(unique(data0_sibling$sibling_id))

length(unique(sibling$sibling_all_id))
```

```
#list the sibling id that does not have individual sibling variables

setdiff(sibling$sibling_all_id, data0_sibling$sibling_id)


table(data0_sibling$sibling_on_roll,useNA="always")


temp <- data0_sibling %>% group_by(id,sibling_id) %>% filter(n() > 1)

nrow(temp)

length(unique(temp$sibling_id))


#remove duplicate cases

nrow(data0_sibling)

data0_sibling <- subset(data0_sibling, select=c("id","sibling_id","sibling_on_roll"))

data0_sibling <- data0_sibling %>% distinct()

nrow(data0_sibling)


temp <- data0_sibling %>% group_by(id,sibling_id) %>% filter(n() > 1)

nrow(temp)


#getting the sibling_on_roll variable for those siblings that do not exist in student data

nrow(sibling_school_NA)

sibling_NA_check <- merge(sibling_school_NA, data0_sibling, by.x=c("id","sibling_all_id"),
by.y=c("id","sibling_id"),all.x=TRUE)

nrow(sibling_NA_check)


#checking the on-roll status for those siblings

table(sibling_NA_check$sibling_on_roll,useNA="always")

sibling_NA_check[is.na(sibling_NA_check$sibling_on_roll)==T,]$sibling_all_id


write.csv(sibling_NA_check, "sibling_NA_check.csv")
```

```

2782 of those 6170 unique siblings do not exist in the raw data including all school years, meaning that 3388 of them exist in the data but not in year 7-11.

```{r, label="Checking if siblings are in the raw data including all years"}

temp <- subset(sibling_school_NA, select=c("sibling_all_id","school_sibl"))

temp <- temp %>% distinct()

temp <- subset(temp, is.na(temp$sibling_all_id)==FALSE)

nrow(temp)

temp <- merge(temp, data0, by.x="sibling_all_id", by.y="id", all.x=TRUE)

nrow(temp)


sum(is.na(temp$school))


```

There are 88420 families identified by the sibling variable in the student data, including 111,883 students.

```{r, label="Creating family id and map of student id to family id"}

#adding the student to the family id vector

data <- data |> mutate(family_all_id=map2(id,sibling_all_id,c))


#sort the family_all_id vector alphabetically

data$family_all_id <- lapply(data$family_all_id, sort)

data_family <- subset(data, select=c("family_all_id"))


#remove repeated entries (of the same family)

nrow(data_family)

data_family <- data_family %>% distinct()

nrow(data_family)


#create family id

```
data_family <- mutate(data_family, family_id = paste0("FAM",row_number()))

#reshape the data from wide to long - each individual id with a family id

family_map <- unnest(data_family, family_all_id)

names(family_map)[names(family_map) == "family_all_id"] <- "id"

str(family_map)

#save a copy of family_map

family_map0 <- family_map


```

Check if the same id appear in two families - each id should only belong to one family: currently 1046 students are associated with more than one family (2730 families in total). Some of them are associated with five families.

```{r, label="List student ids associated with multiple families"}

family_duplicates <- family_map %>% group_by(id) %>% filter(n() > 1)

family_duplicates <- merge(family_duplicates, data_family, by=c("family_id"))

nrow(family_duplicates)

length(unique(family_duplicates$id))


family_duplicates_freq <- as.data.frame(table(family_duplicates$id))

table(family_duplicates_freq$Freq)

```

In 2705 out of 2730 cases, the other families a student id is associated with are a subset of the largest family the student is associated with.

```{r, label="Check if multiple families associated with the same id are subset of each other"}

#get the family for each id that has the largest number of individuals

family_longest <- family_duplicates %>%

  group_by(id) %>%

  slice(which.max(lengths(family_all_id)))


```

```
names(family_longest)[names(family_longest) == "family_all_id"] <- "family_all_id_longest"

names(family_longest)[names(family_longest) == "family_id"] <- "family_id_longest"


#merge the longest family for each id to each row of family associated with the id

family_duplicates <- merge(family_duplicates, family_longest, by=c("id"), all.x=TRUE)


#check if all the other families are a subset of the largest one - members of the other families are members of the largest family

set_diff <- function(list1, list2) {

  list1[!list1 %in% list2]

}


family_duplicates$diff  <- mapply(set_diff, family_duplicates$family_all_id, family_duplicates$family_all_id_longest)


#list the number of members that do not belong to the largest families

family_duplicates$diff_length  <- lapply(family_duplicates$diff, length)

table(as.numeric(family_duplicates$diff_length))

```

For the 2705 cases (associated with 1046 student id), use the family id of the largest family to replace the multiple smaller families in the original family map.

```{r, label="Replace the other families with the largest families for the cases where they are subset of largest families"}

#get the subset of cases where the other families are a subset of the largest family

family_duplicates_nonproblematic <- subset(family_duplicates, family_duplicates$diff=="character(0)")

nrow(family_duplicates_nonproblematic)


#create the family map for these IDs using the family id of the largest family

family_duplicates_nonproblematic <- subset(family_duplicates_nonproblematic, select=c("id","family_id_longest"))

family_duplicates_nonproblematic <- family_duplicates_nonproblematic %>% distinct()

nrow(family_duplicates_nonproblematic)
```

#check in the new map, each id is associated with only one family id

family_nonproblematic_duplicates <- family_duplicates_nonproblematic %>% group_by(id) %>% filter(n() > 1)

nrow(family_nonproblematic_duplicates)

#replacing the family id with the family id of the largest family in the original family map for these IDs

nrow(family_map)

family_map <- merge(family_map, family_duplicates_nonproblematic, by="id", all.x=TRUE)

nrow(family_map)

family_map$family_id <- ifelse(is.na(family_map$family_id_longest)==FALSE & family_map$family_id_longest!= family_map$family_id, family_map$family_id_longest, family_map$family_id)

family_map <- subset(family_map, select=c("id","family_id"))

family_map <- family_map %>% distinct()

nrow(family_map)

```

There are 25 cases where the largest family associated with an id does not include a member that exists in another family the same student id is associated with. In such cases, we merge the two families to be one new family. 13 unique merged families are created.

```{r, label="Merge the families that associate with the same student id but are not subset of each other"}

#list such problematic cases

family_duplicates_problematic <- subset(family_duplicates, family_duplicates$diff!="character(0)")

nrow(family_duplicates_problematic)

temp <- as.data.frame(table(family_duplicates_problematic$id))

table(temp$Freq)

#output such cases to csv file

temp<-family_duplicates_problematic

```
temp$family_all_id<-as.character(temp$family_all_id)

temp$family_all_id_longest<-as.character(temp$family_all_id_longest)

temp$diff<-as.character(temp$diff)

temp <- merge(temp, subset(data,select=c("id","sibling_all_id")),

                                by.x="diff",by.y="id",all.x=TRUE)

temp$sibling_all_id <- as.character(temp$sibling_all_id)

temp$diff_length <- as.numeric(temp$diff_length)

write.csv(temp, "family_problematic.csv")

#for the problematic cases, merge families into one

family_duplicates_problematic <- family_duplicates_problematic |>
mutate(family_all_id_merge=map2(diff,family_all_id_longest,c))

#check the original families are subsets of the merged families

family_duplicates_problematic$diff3 <- mapply(set_diff, family_duplicates_problematic$family_all_id,
family_duplicates_problematic$family_all_id_merge)

family_duplicates_problematic$diff4 <- mapply(set_diff,
family_duplicates_problematic$family_all_id_longest,
family_duplicates_problematic$family_all_id_merge)

family_duplicates_problematic2 <- subset(family_duplicates_problematic,
family_duplicates_problematic$diff3!="character(0)" |
family_duplicates_problematic$diff4!="character(0)")

nrow(family_duplicates_problematic2)

#check the merged families are unique for each id

family_duplicates_problematic$family_all_id_merge <-
lapply(family_duplicates_problematic$family_all_id_merge, sort)

data_family_merge <- subset(family_duplicates_problematic, select=c("id","family_all_id_merge"))

nrow(data_family_merge)

data_family_merge <- data_family_merge %>% distinct()

nrow(data_family_merge)

data_family_merge_duplicates <- data_family_merge %>% group_by(id) %>% filter(n() > 1)
```

```
nrow(data_family_merge_duplicates)


#create family id for the merged families

data_family_merge <- subset(data_family_merge, select=c("family_all_id_merge"))

nrow(data_family_merge)

data_family_merge <- data_family_merge %>% distinct()

nrow(data_family_merge)

data_family_merge <- mutate(data_family_merge, family_id_merge =
paste0("FAM_C",row_number()))


#create family map for the members of the merged families

family_map_merge <- unnest(data_family_merge, family_all_id_merge)

names(family_map_merge)[names(family_map_merge) == "family_all_id_merge"] <- "id"


#check there are not duplicate members in the merged families

family_merge_duplicates <- family_map_merge %>% group_by(id) %>% filter(n() > 1)

nrow(family_merge_duplicates)
```

We've created a fmaily map with each of the 104029 students associated with a unique family ID.

```{r, label="Replace the family id with the family id of the merged family for those in the merged families"}

#for those IDs in the merged families, replace the family id using the family id of the merged family

nrow(family_map)

family_map <- merge(family_map, family_map_merge, by="id", all.x=TRUE)

nrow(family_map)

family_map$family_id <- ifelse(is.na(family_map$family_id_merge)==FALSE,
family_map$family_id_merge, family_map$family_id)


family_map <- subset(family_map, select=c("id","family_id"))

family_map <- family_map %>% distinct()

nrow(family_map)
```

#check each id is associated with just one family id in the updated family map

family_duplicates2 <- family_map %>% group_by(id) %>% filter(n() > 1)

nrow(family_duplicates2)

#remove students not in the student data

temp <- subset(data,select=c("id"))

family_map <- merge(family_map, temp, by="id")

nrow(family_map)

colSums(is.na(family_map))

```

There are 87909 families in total. 72887 with only one family member. The rest have 2 to 5 family members.

```{r, label="Check final family size"}

length(unique(family_map$family_id))

#check size of families

family_freq <- as.data.frame(table(family_map$family_id))

table(family_freq$Freq)

table(data$sibling_num)

```

Number of families per school ranges from 198 to 1510, with median of 769.5. Average number of pupils per family in a school varies from 1.01 to 1.30.

```{r, label="Check number of families and family size per school"}

#Get information for each family member

temp <- subset(data,select=c("id","school","EAL","FSM"))

family_data <- merge(temp, family_map, by=c("id"), all = TRUE)

str(family_data)

#check if there is a student without family id

sum(is.na(family_data$family_id))

```
#check if there is a family member without student information

sum(is.na(family_data$id))

#look at the distribution of the number of families across schools

family_school <- family_data %>%

  group_by(school) %>% summarise(family_number = n_distinct(family_id))

summary(family_school$family_number)

#the ratio of students to families (i.e., the average pupils per family) across schools - only include family members with student information

family_size_school <- aggregate(id~family_id+school, data = family_data, n_distinct)

family_size_school <- aggregate(id~school, data = family_size_school, mean)

summary(family_size_school$id)
```

Out of the 87909 families, there are 436 families with inconsistent FSM status, and 3311 families have inconsistent EAL status. No families have members belong to multiple schools.

```{r, label="Check families with inconsistent EAL and FSM status"}

#check if all family members belong to the same school, and have different EAL/FSM status

family_inconsistent <- family_data %>%

  group_by(family_id) %>%

  summarise(distinct_school = n_distinct(school),

            distinct_EAL = n_distinct(EAL),

            distinct_FSM = n_distinct(FSM))


table(family_inconsistent$distinct_school, useNA="always")

table(family_inconsistent$distinct_EAL, useNA="always")

table(family_inconsistent$distinct_FSM, useNA="always")


```

Out of the 65,587 families with at least one member that does have non-missing EAL (55782 of these families have only one member), 755 have inconsistent EAL values.

```{r, label="Check families with inconsistent EAL (excluding missing)"}

family_inconsistent_EAL <- subset(family_data, is.na(family_data$EAL)==FALSE)

temp <- as.data.frame((table(family_inconsistent_EAL$family_id)))

table(temp$Freq)


family_inconsistent_EAL <- family_inconsistent_EAL %>%

  group_by(family_id) %>%

  summarise(distinct_EAL = n_distinct(EAL))


table(family_inconsistent_EAL$distinct_EAL, useNA="always")

```

There are 22322 families with missing EAL for all family members. Family-level EAL and FSM is created using the rule that as long as at least one member of the family has EAL/FSM status, the family is treated as EAL/FSM = 1, otherwise 0. There are 14848 families with EAL=1, and 25455 families with FSM=1.

```{r, label="Creating family level EAL and FSM"}

#check number of families with all members with missing EAL

table(family_data$EAL, useNA="always")

family_data$EAL1 <- ifelse(is.na(family_data$EAL),-1,family_data$EAL)

temp <- aggregate(EAL1 ~ family_id+school, data = family_data, mean)

table(temp$EAL1, useNA="always")


#replace missing EAL with 0

family_data$EAL0 <- family_data$EAL

family_data$EAL <- ifelse(is.na(family_data$EAL),0,family_data$EAL)


#generating the sum of EAL and FSM for all members in a family

data_strtf <- aggregate(cbind(EAL,FSM) ~ family_id+school, data = family_data, sum)


table(data_strtf$EAL,useNA="always")

table(data_strtf$FSM,useNA="always")
```

#family is coded as FSM/EAL as long as at least one member is coded as FSM/EAL

data_strtf$eal <- if_else(data_strtf$EAL>0, 1, 0)

data_strtf$fsm <- if_else(data_strtf$FSM>0, 1, 0)

table(data_strtf$eal,useNA="always")

table(data_strtf$fsm,useNA="always")

```

Size of strata varies from 1 to 1150.

```{r, label="Randomisation"}

set.seed(2023)

#creating blocks for the two stratification variables within school

data_strtf$block <- with(data_strtf, paste(eal, fsm, school, sep = "_"))

temp <- as.data.frame(table(data_strtf$block))

summary(temp$Freq)

#perform stratified randomisation within school

data_strtf$arm <- block_ra(blocks = data_strtf$block, num_arms = 2, prob_each = c(.5, .5))

#show number of families in each arm in each strata

table(data_strtf$block,data_strtf$arm)

```

There are 52007 students assigned to arm1, and 52022 assigned to arm2.

```{r, label="Check randomisation results"}

#assign the randomisation results to all family members in the student data

student_arm <- merge(family_data, subset(data_strtf,select=c("family_id","arm", "eal","fsm")),
by="family_id",all.x=TRUE)

```
#check if there is a student without an assigned arm

sum(is.na(student_arm$arm))

#check number of students per arm by school, EAL and FSM

table1(~as.factor(school)+as.factor(EAL)+as.factor(FSM)|arm, data=student_arm)


```


Check the possible values for the SEN category variables.

```{r, label="Check Sen status"}

#getting the variables for individual siblings in student data

data_sen <- data %>%

   select(id, starts_with('sen_')) %>%

   select(id, ends_with('_category_code') | ends_with('_category_description') )


#reshape to long form

data_sen <- data.table::melt(setDT(data_sen),id="id",variable.name="sen",measure=patterns(

                sen_category_code=".*_category_code",

                sen_category_description=".*_category_description"))


#check the possible values

table(data_sen$sen_category_code, useNA="always")

table(data_sen$sen_category_description, useNA="always")


```


Create a binary SEN status = 1 if SEN category is not missing and not "NO Specialist Assessment". Sen status = 0 otherwise.

```{r, label="Creating binary Sen status for each student id"}

data_sen$sen_indicator <- ifelse(is.na(data_sen$sen_category_description)==TRUE | data_sen$sen_category_description == "No Specialist Assessment", 0, 1)
```

```
data_sen_id <- aggregate(sen_indicator ~ id, data=data_sen, sum)
table(data_sen_id$sen_indicator, useNA="always")

data_sen_id$sen <- ifelse(data_sen_id$sen_indicator >0, 1, 0)
table(data_sen_id$sen, useNA="always")
```

```{r, label="Clean randomisation output file"}
#remove the additional EAL variables
student_arm <- student_arm %>% select(-EAL, -EAL1)

#rename EAL and FSM
colnames(student_arm)[colnames(student_arm) == "EAL0"] <- "student_EAL"
colnames(student_arm)[colnames(student_arm) == "FSM"] <- "student_FSM"
colnames(student_arm)[colnames(student_arm) == "eal"] <- "family_EAL"
colnames(student_arm)[colnames(student_arm) == "fsm"] <- "family_FSM"

#rename T1 and T2 to be "Control" and "Intervention"
table(student_arm$arm, useNA="always")
student_arm$arm <- ifelse(student_arm$arm=="T2", "Intervention", "Control")

#adding school urn
temp <- subset(school_confirm, select=c("region.identifiers.urn","name"))
nrow(student_arm)
student_arm <- merge(student_arm, temp, by.x="school", by.y = "name")
nrow(student_arm)
colnames(student_arm)[colnames(student_arm) == "region.identifiers.urn"] <- "school_URN"

#adding SEN status variable
temp <- subset(data_sen_id, select=c("id","sen"))
nrow(student_arm)
```

```
student_arm_final <- merge(student_arm, temp, by="id")

nrow(student_arm_final)


#adding ethnicity and care variables

temp <- subset(data, select=c("id","extended_details.data.ethnicity",
"extended_details.data.in_lea_care", "extended_details.data.ever_in_care"))

nrow(student_arm_final)

student_arm_final <- merge(student_arm_final, temp, by="id")

nrow(student_arm_final)


colnames(student_arm_final)[colnames(student_arm_final) == "extended_details.data.ethnicity"] <-
"student_ethnicity"

colnames(student_arm_final)[colnames(student_arm_final) == "sen"] <- "student_sen"

colnames(student_arm_final)[colnames(student_arm_final) == "extended_details.data.in_lea_care"]
<- "student_in_lea_care"

colnames(student_arm_final)[colnames(student_arm_final) == "extended_details.data.ever_in_care"]
<- "student_ever_in_care"



str(student_arm_final)

#save randomisation result

write.csv(student_arm_final, "student_arm_final.csv")

```


Getting counts for treatment groups

```{r, label="Summary statistics of randomisation results"}

student_treatment <- subset(student_arm_final, student_arm_final$arm=="Intervention")


#Total number of young people  (number)

length(unique(student_treatment$id))


# Number of Black young people            (number)
```

```
# Number of Asian young people            (number)

# Number of mixed or multiple ethnicity young people      (number)

# Number of Gypsy Romany Traveller Young People       (number)

# Number of other minority ethnicity (not included above) young people    (number)

temp <- as.data.frame(table(student_treatment$student_ethnicity, useNA="always"))

sum(temp$Freq)

# write.csv(temp, "ethnicity_treatment.csv")


#Read the ethnicity categorisation file

ethnicity_category<-read_excel("ethnicity_treatment_categorised_v4.xlsx")

ethnicity_category<-subset(ethnicity_category, select=c("Code","Label"))

ethnicity_final<-merge(ethnicity_category, temp, by.x="Code", by.y= "Var1", all=TRUE)

ethnicity_final$Freq <- ifelse(is.na(ethnicity_final$Freq)==TRUE, 0, ethnicity_final$Freq)

sum(ethnicity_final$Freq)


subset(ethnicity_final,is.na(ethnicity_final$Label)==TRUE)


ethnicity_final$Label <- ifelse(ethnicity_final$Code=="Libyan", "Number of other minority ethnicity (not
included above) young people", ethnicity_final$Label)

ethnicity_final$Label <- ifelse(is.na(ethnicity_final$Code)==TRUE, "NA (not reported)",
ethnicity_final$Label)


ethnicity_final <- aggregate(Freq ~ Label, data=ethnicity_final, sum)

ethnicity_final

sum(ethnicity_final$Freq)


#Number of Looked After Young People (number)

table(student_treatment$student_in_lea_care, useNA="always")

#table(student_treatment$extended_details.data.ever_in_care, useNA="always")


# Number of Young People with SEND   (number)
```

```
table(student_treatment$student_sen, useNA="always")

# Number of FSM-eligible young people
table(student_treatment$student_FSM, useNA="always")

# Number of EAL young people  (number)
table(student_treatment$student_EAL, useNA="always")
```

```{r, label="Randomisation output file to be shared"}
#keep necessary columns and add upi
temp <- subset(data, select=c("id","upi"))
student_arm_share <- student_arm_final %>% select(id,family_id,arm, school_URN)
student_arm_share <- merge(student_arm_share, temp, by="id")

str(student_arm_share)
length(unique(student_arm_share$family_id))
length(unique(student_arm_share$school_URN))
length(unique(student_arm_share$upi))

table(student_arm_share$arm)
#save randomisation result
write.csv(student_arm_share, "262400452 BITUP randomisation_KP shared with BIT 20.10.2023.csv")
```