STATISTICAL ANALYSIS PLAN

Grassroots Efficacy Trial:

Evaluation of an anti-conflict intervention by cluster randomised controlled trial

IOE – UCL's Faculty of Education and Society

Principal investigator: Becky Taylor

YOUTH ENDOWMENT FUND

Grassroots Efficacy Trial: Evaluation of an anti-conflict intervention by cluster randomised controlled trial

Statistical analysis plan

Evaluating institution: IOE – UCL's Faculty of Education and Society

Principal investigator(s): Becky Taylor

YEF statistical analysis plan

Project title	Grassroots Efficacy Trial: Evaluation of an anti-conflict intervention by cluster randomised controlled trial
Developer (Institution)	Behavioural Insights Team (BIT)
Evaluator (Institution)	IOE – UCL's Faculty of Education and Society
Principal investigator(s)	Becky Taylor
SAP author(s)	Sal Riordan, Jeremy Hodgen, Becky Taylor
Trial design	Two-armed cluster randomised controlled trial with random allocation at the school level
Trial type	Efficacy
Evaluation setting	School
Target group	11- to 14-year-olds (school years 7 to 9) in mainstream secondary schools in England and Wales
Pupil age range and key stage	11- to 14-year-olds, KS3

Number of participants	100 schools, 60 000 participants
Primary outcome and data source	School attendance (NPD and SAIL)
Secondary outcome and data source	Strengths and Difficulties Questionnaire (self-report survey) Disciplinary report measures (school data) Bullying and Cyberbullying Scale for Adolescents (self-report survey) Peer Conflict Scale (self-report survey) Feelings of safety at school (self-report survey)

SAP version history

Version	Date	Changes made and reason for revision
1.3	15.7.2024	Minor edits to clarify primary outcome measure and number of schools randomised.
1.2	29.5.2024	Updated detail of disciplinary report data analysis.
1.1	19.1.2024	Addition of effect size calculations for FSM-eligible pupils to reflect YEF's latest SAP template.
1.0 [*original*]	1.12.2023	

Any changes to the design or methods need to be discussed with the YEF Evaluation Manager and the developer team prior to any change(s) being finalised. Describe in the table above any agreed changes made to the evaluation design. Please ensure that these changes are also reflected in the SAP (CONSORT 3b, 6b).

Table of contents

SAP version history 2

Table of contents 3

Introduction 4

Design overview 4

Sample size calculations overview 6

Analysis 9

References 20

Appendix 23

Introduction

Grassroots is an anti-conflict programme that aims to reduce bullying and conflict in schools by empowering socially influential pupils to positively impact their fellow pupils' behaviours. It has evidence from a randomised controlled trial in the USA but has not been delivered previously in England and Wales. Further details of what intervention consists of are presented in the trial protocol, as well as how the intervention is hypothesized to reduce bullying and peer-to-peer conflict, increase pupil attendance, and improve pupils' feelings of safety in school.

Having conducted a pilot trial of Grassroots during the spring and summer terms of 2023, the efficacy trial began in September 2023. The purpose of the efficacy trial is to determine the impact of Grassroots under ideal circumstances when schools are supported by the developer to deliver the intervention.

The efficacy trial includes an impact evaluation, the main objective of which is to determine whether the delivery of Grassroots increases school attendance, in the immediate term and after two years. The secondary objectives of the evaluation are to estimate the immediate impact of the intervention on pupils' bullying and conflict behaviours, as well as pupils' feelings of safety in school. Analysis will be conducted to determine if the impacts of Grassroots differ by ethnicity and free school meal eligibility (FSM). This statistical analysis plan provides the details of how these impacts will be estimated.

Design overview

The trial is designed to answer the primary research question:

I1 Does empowering Key Stage 3 pupils in English and Welsh secondary schools to positively impact fellow pupils' social behaviours increase school attendance (a) at the end of the intervention, and (b) after 2 years, compared with Key Stage 3 pupils in schools using business-as-usual conflict/bullying reduction activities?

The secondary research questions for the trial are:

I2 Does empowering Key Stage 3 pupils in English and Welsh secondary schools to positively impact fellow pupils' social behaviours improve social and emotional outcomes as measured by the Strengths and Difficulties Questionnaire, compared with Key Stage 3 pupils in schools using business-as-usual conflict/bullying reduction activities?

I3. Does empowering Key Stage 3 pupils in English and Welsh secondary schools to positively impact fellow pupils' social behaviours reduce conflict between pupils as measured by (a) school disciplinary reports; and (b) the Peer Conflict Survey

compared with Key Stage 3 pupils in schools using business-as-usual conflict/bullying reduction activities?

I4 Does empowering Key Stage 3 pupils in English and Welsh secondary schools to positively impact fellow pupils' social behaviours reduce bullying perpetration and victimisation as measured by the Bullying and Cyberbullying Scale for Adolescents compared with Key Stage 3 pupils in schools using business-as-usual conflict/bullying reduction activities?

I5 Does the impact of Grassroots differ by ethnicity or free school meal (FSM) eligibility?

A summary of the trial design is presented in Table 1. These details are in line with version 1.0 of the trial protocol, which provides more information regarding the design of the trial.

Table 1. Trial design

Trial design, including number of arms		Two-arm, cluster randomised efficacy trial
Unit of randomisation		School (1:1 allocation ratio)
Stratification variables (if applicable)		Recruitment region
Primary outcome	variable	Attendance in Summer term 2023-2024 Follow-up: attendance in Autumn term 2025-26
	measure (instrument, scale, source)	% school sessions attended in Summer term 2023-24, NPD (England), SAIL (Wales) Follow up: % school sessions attended in Autumn term 2025-26, NPD, SAIL
Secondary outcome(s)	variable(s)	1. Emotional problems, conduct problems, hyperactivity, peer relationship problems and pro-social behaviour 2. Teacher-reported peer-to-peer conflictBullying perpetration and victimisation 3. Student-reported peer-to-peer conflict 4. Feeling safe in school
	measure(s) (instrument, scale, source)	1. Self-report survey (SDQ) June/July 2024 2. School disciplinary reports from school management systems, Summer term 2023-24 3. Bullying and Cyberbullying Survey (BCS-A), self-report survey June/July 2024 4. Peer Conflict Survey (short version), self-report survey June/July 2024

		5. Feeling safe in school, self-report survey June/July 2024
Baseline for primary outcome	**variable**	Attendance in Autumn term, 2023-24
	measure (instrument, scale, source)	% school sessions attended in Autumn term 1, 2023, NPD (England), SAIL (Wales)
Baseline for secondary outcome	**variable**	1. Emotional problems, conduct problems, hyperactivity, peer relationship problems and pro-social behaviour 2. Teacher-reported peer-to-peer conflictBullying perpetration and victimisation 3. Student-reported peer-to-peer conflict 4. Feeling safe in school
	measure (instrument, scale, source)	1. Self-report survey (SDQ), October 2023 2. School disciplinary reports from school management systems, Autumn term 2023-24 3. 4. Bullying and Cyberbullying Survey (BCS-A), self-report survey October 2023 5. Peer Conflict Survey (short version), self-report survey October 2023 6. Feeling safe in school, self-report survey October 2023.

Sample size calculations overview

Please ensure all details are in line with the latest version of the protocol.

Table 2. Summary of sample size calculations

		Protocol		**Randomisation**	
		Overall	FSM	Overall	FSM
Minimum Detectable Effect Size (MDES)		0.2	0.2	0.19	0.19
	level 1 (participant)	0.5	0.5	0.5	0.5

		Protocol		Randomisation	
		Overall	FSM	Overall	FSM
Pre-test/ post-test correlations	level 2 (cluster)	0.5	0.5	0.5	0.5
Intracluster correlations (ICCs)	level 1 (participant)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		540	130	571	161
Number of clusters	Intervention	50	50	53	53
	Control	50	50	53	53
	Total	100	100	106	106
Number of participants	Intervention	27,000	6,500	28,818	7,742
	Control	27,000	6,500	31,662	9,340
	Total	54,000	13,000	60,480	17,082

Sample size assumptions

The sample size was determined a priori using the mrss.cra2 function from the PowerUpR package in R. For the assumptions listed in Table 2, for a 2-level model with treatment at level 2, including the standard requirements of 0.8 for power and 0.05 for alpha, we determined a

minimum sample size of 91 schools for our main scenario. This sample size estimation is for our primary outcome (attendance), and our primary population of interest is all KS3 pupils.

FSM assumptions. Using the estimate that 23.8% of pupils are eligible for FSM (from government data for England's secondary schools in 2022-23), we estimated the numbers of FSM-eligible pupils in the trial (shown in Table 2). We then determined a minimum sample size of 94 schools for this subgroup of pupils. *Correlation assumptions.* We assumed a standard value of 0.15 for the intracluster correlation coefficient (ICC). ICCs between 0.1 and 0.2 are commonly assumed for education trials, supported by past experience of evaluations by the Education Endowment Foundation (Demark, 2019). Historically, however, there is more information regarding ICCs related to examination outcomes than attendance. We will calculate and publish the ICC for the primary outcome of this trial to support future statistical modelling of evaluations measuring attendance.A recent study has shown that studies archived in the EEF and NPD databases have median ICCs of between 0.10 and 0.13 for KS2 and KS4 interventions (Singh et al, 2023). We have used a more conservative estimate of 0.15. Similarly, we have taken standard (but cautious) values for pre-/post-test correlations and conducted sensitivity analysis to check that sample sizes are not reliant on unrealistic values for these variables.

Minimum Detectable Effect Size. Outcome measures for previous evaluations of the Grassroots intervention (previously called ROOTS) have included disciplinary reports of pupil conflict, talking to friends about how to reduce conflict, wearing anti-conflict wristbands, and social norms (Paluck et al., 2016); and measures of bullying perpetration and victimisation, social norms and school climate (Bowes et al., 2019). However, it is not possible to calculate effect sizes from the information in the published articles. Our power calculations are therefore based on a MDES of 0.2 for the primary outcome measure (attendance), in line with standard practice for randomised controlled trials in education (Hutchison and Styles, 2010).

Further modelling. We have also run calculations to ensure that the sample size is sufficiently powered to detect a MDES of 0.2 on the secondary outcome measures. For cost reasons, we will only collect SDQ data from a random sample of one third of KS3 pupils. We have therefore performed sample size estimations for this reduced number of pupils per school. Taking a cautious approach to estimating sample size, we have also assumed smaller pre-test/post-test correlations for secondary measures (of 0.4), supported by test-retest reliability measurements recorded for the Bullying and Cyberbullying Survey (Özbey, H & Öznur Başdaş, Ö. (2020). We have also carried over a conservative estimate of ICC=0.15 (expecting it to be lower for SDQ). As a result, we have estimated that we need 97 schools to detect an MDES of 0.2 on all our measures and the developers aimed to recruit 115 schools before July 2023, allowing for some attrition.

Randomisation

106 schools were recruited to the intervention (42 in Greater London, 16 in Wales, 6 in the West of England, 17 in the West Midlands, and 25 in North-West England). Randomisation was conducted in October 2023. It was carried out on a 1:1 ratio of allocation to treatment and control at school level, stratified by the five recruitment regions. Stratification was to ensure the distribution of treatment schools between regions to facilitate research assistant recruitment and programme delivery by the developers. In two of the recruitment regions (South West England and Wales) the same research assistant will deliver the intervention, but we applied to randomisation procedure to divide the schools in both these regions equally into the control and intervention group to ensure that the trial randomisation was stratified by country (Wales/England).

Using a predetermined seed for replicability, each school was assigned a random number between 0 and 1. The schools were ordered by this random number allocation and those in the first half of each region were assigned to the control group, those in the second half assigned to the intervention group. The randomisation procedure was conducted and saved in R, including the random seed (the syntax is given in the appendix). As a result, 53 schools (with 31,662 KS3 pupils) were assigned to the control group and 53 schools (with 28,818 KS3 pupils) to the intervention group. These updated figures were used in our updated minimum detectable effect size (MDES) calculations. After randomisation, we therefore have an estimated MDES for our primary outcome of 0.19, as detailed in Table 2. Post randomisation, it was found that there were more FSM-eligible pupils than expected, and that the trial was also powered to detect effect sizes of 0.19 in this subgroup.

Analysis

We have taken the approach to use multi-level modelling to estimate the impact of the Grassroots intervention on pupils' attendance, pupils' bullying and conflict behaviours, and pupils' feelings of safety in school. The purpose of the using a hierarchical model is to account for clustering at the school level. The Grassroots intervention brings together children from different classes and across different year groups, and therefore it is not clearly clustered by class or year group. We decided not to unnecessarily complicate the model with these levels. It was also decided not to use levels to account for clustering by region because we have no theoretical reason for considering regional differences, and because there are only 5 recruitment regions, insufficient for multi-level modelling (Bell et al., 2014; Kreft, & de Leeuw, 1998).

The primary outcome for the evaluation will be a measure of the impact of the evaluation in English and Welsh schools. All other analyses are exploratory. It is a core principle of our funder to explore what works for whom. In particular, we will be exploring the broad patterns by which Grassroots impacts students from ethnic minorities and those eligible for free school meals, because these groups are known to have lower school outcomes, including GSCE

results (Francis-Devine et al., 2023) and exclusions (Timpson, 2019). To fully describe broad patterns, we are using several measures with multiple subgroups. As a result, there is a higher risk of false positives overall. Some of these analyses will be underpowered, increasing the chance of false negatives in some cases. All exploratory analyses must therefore be treated with caution and not interpreted individually. Our approach is to divide these exploratory analyses into two groups. For outcomes with the highest expectation of identifying the impact of Grassroots (six secondary outcomes from survey instruments), as well as those with the highest priority to explore structural inequalities (attendance in England and Wales by ethnicity and FSM eligibility), we will run multi-level regression modelling. We will report an updated significance-level using a familywise error correction to these 18 tests to support careful interpretation of these data. The remaining 45 secondary outcomes (see below) will be reported as descriptive statistics only.The analysis will be conducted in R (version 4.3.1). Indicative syntax for the main elements of the analysis is provided in the appendix. The general methods of analysis were decided prior to the data collection for the main trial. The survey measures (SDQ, BCS-A, peer conflict and norms) were selected after the pilot study tested a variety of measures of conflict and bullying behaviours in the summer of 2023, including the Olweus measure of bullying and perpetration. The disciplinary report measure was tested using a sample of disciplinary report data collected from trial schools in October 2023.

Primary outcome analysis

The primary outcome for the evaluation is attendance. The impact on attendance will be estimated using a random-intercept model given by Equation 1. We take an ANCOVA approach, using baseline attendance as a covariate (instead of analysing change in attendance as an outcome), because of the increase of power this gives in randomised studies (van Breukelen, 2013).

Equation 1. Primary outcome model

$$y_{ij} = \beta_0 + \beta_1 PreTest_{ij} + \beta_2 Treat_j + \mu_j + e_{ij}$$

where

y_{ij}	is the post-intervention attendance of *i*th pupil at school *j*
β_0	is the grand intercept (average post-attendance of non-intervention schools for pre-attendance of zero)
$PreTest_{ij}$	is the pre-intervention attendance of *i*th pupil at school *j* (with associated regression coefficient β_1)

$Treat_j$ is a dummy variable = 1 if school *j* is intervention school (with associated regression coefficient β_2)

μ_j is the school-effect (for school *j*, the intercept is $\beta_0 + \mu_j$) and it is assumed that $u_j \sim N(0, \sigma_u^2)$

e_{ij} is the residual for *i*th pupil at school *j* and it is assumed that $e_{ij} \sim N(0, \sigma_e^2)$

The attendance variables y_{ij} and $PreTest_{ij}$ are the percentage of school sessions attended in the Summer and Autumn terms of the 2023-24 school year respectively. A school session refers to either a morning or afternoon when the school is open and pupils are expected to attend. These values will be calculated from the number of school sessions missed and the total number of school sessions, which will be taken from the National Pupil Database (NPD) for English schools and the SAIL databank for Welsh schools. Because we are unable to transfer pupil data between these two databases, we will run the statistical model for pupils in English schools and pupils in Welsh schools separately. We will report the results from both analyses separately, and also report a weighted average of the effect sizes using meta-analysis. This averaging will account for the larger number of English schools (and the greater precision in the finding for England) in comparison to Welsh schools.

We are keeping to a simple model with one covariate for the primary analysis to increase transparency and aid interpretation. This will be reported for both English and Welsh schools and will provide an estimate of the impact of the intervention overall. Further subgroup analysis will be conducted as part of exploratory tests (described in the subgroup section below). Corvariates will be added to the primary model to analyse attendance for subgroups of pupils (FSM eligibility and ethnicity). This will be conducted separated for Welsh and English attendance data. We will use the five categories of ethnicity from the YEF's policy on collection of demographic data (2023).

We will verify that the assumptions for linear modelling are met. If not, we will instead fit the data using a transformation or generalised linear model, depending on the data.

Secondary outcome analysis

The same model specification to the primary analysis model of Equation 1 will be used to analyse the impact of Grassroots on 6 secondary outcome measures, but with the dependent variable y_{ij} (and associated baseline covariate $PreTest_{ij}$) replaced with measures of bullying and peer conflict from our pupil surveys. The model will be applied to the four subscales of the Bullying and Victimisation survey, as well as two from the peer conflict scale. For these outcomes, we will analyse data from pupils in English and Welsh schools together.

Bullying and Cyberbullying Scale for Adolescents (BCS-A). This BCS-A scale measures victimisation and perpetration experiences of young people (Thomas et al., 2018). It has four subscales, each with 20 items, measuring bullying experiences, cyberbullying experiences, perpetration experiences, and cyber-perpetration experiences. These four measures will each be analysed using the model of Equation 2.

Peer Conflict Scale (PCS) The PCS (Youth Version) is a questionnaire designed to measure the aggressive behaviours of young people (Marsee et al., 2014). It measures physical proactive and active aggression, as well as relational proactive and relational active aggression. We are using the shortened, 20-item version to consider the impact of Grassroots on aggressive behaviours between peers using the model of Equation 2 (Pechorro et al, 2012). We will use two subscales of the shortened version (measuring physical aggression and relational aggression).

Error correction

We will be applying a familywise error correction using the Benjamini-Hochberg (BH) procedure. Instead of using a cut-off of 0.05 for statistical significance, we will apply the BH procedure to the collection of *p*-values that result from our statistical modelling, to determine which are significant. The BH procedure effectively reduces the level of statistical significance. The purpose of doing this is to reduce the probability of false positives because altogether we are testing 18 hypotheses (regarding the impact of Grassroots on a variety of measures and the impact of several pupil and school characteristics). We will treat the collection of secondary tests and most subgroup tests as one family and exploratory. The primary outcome will not be included in this family because it is not exploratory in the same way and to enable to primary trial evaluation result to be comparable to other trials. s.

Further data reporting

In addition to the multi-level modelling of 6secondary measures, we will be collecting and reporting on a further set of secondary measures to enable comparisons to be made with the US study and to provide data for further research and analysis. We will report a further 8 measures in this way, but will not run the statistical modelling described in Equation 1. The descriptive data will provide support for identifying broad patterns in the findings. Because of the multiple tests, as well as the underpowering of subgroup analyses, care must be taken in interpreting this data. No single data point will be interpreted as showing a significant impact of the intervention in itself.

The outcomes we will report on in this way are responses to the Strengths and Difficulties Questionnaire (SDQ), two norm measures and disciplinary report data, and feelings of safety in school.

Strengths and Difficulties Questionnaire (SDQ). The SDQ is a standard instrument for assessing young people's mental health. It consists of 25 items, providing measuring emotional symptoms, conduct problems, hyperactivity, peer relationship problems and prosocial behaviour. We will report on the 5 subscales of the SDQ in our control and intervention arms of the trial, providing a breakdown for ethnicity and FSM eligibility.

Feelings of safety in school. Our survey includes two questions regarding pupils' feelings of safety in school. The first is a question from Ofsted's pupil questionnaire (I feel safe when I am at school, All the time/Most of the time/Some of the time/Almost never/Never). The second (During past 30 days, did you miss school due to feeling unsafe?) has been used in studies of absenteeism and pupil behaviours (Epstein et al., 2020). We will report on the responses to these questions, including breakdowns by ethnicity and FSM eligibility. The primary reason for collecting this data, however, is so that we can conduct sensitivity analysis on our primary outcome. We will analyse the attendance data for smaller, more targeted, groups of the pupil population, to consider the impact of Grassroots on students who do not feel safe at school and those who report missing school in the last 30 days due to feeling unsafe.

Norms and disciplinary report measures. We are using two survey measures (descriptive and normative norms) that were used in the US study of Grassroots (Paluck et al., 2016). We will report these survey results in order that we can make a direct comparison with the US evaluation (which used descriptive statistics and not multi-level modelling). The US evaluation found a drop in the number of disciplinary reports made at schools in the intervention. In the UK, schools use management information systems (MIS) to record disciplinary incidents. Because school practices vary widely, the pilot study was too small to determine whether it was both practical and effective to use this data for the evaluation. MIS contain large numbers of reports that are irrelevant to the intervention. We explored the use of machine coding using a sample of disciplinary reports collected from trial schools in October 2023. This demonstrated that we were unable to identify peer-to-peer conflict in disciplinary reports with sufficient precision and reliability to use this as a measure. T However, we are working with the developers to hand-code textual disciplinary reports. If we are able to do this, we will report the differences in counts of peer-to-peer disciplinary reports between control and intervention schools. The measure is underpowered, because not all trial schools provide disciplinary reports in text form and should therefore be interpreted with caution.

Subgroup analyses

The evaluation will analyse the impact of Grassroots on the primary outcome (attendance) for pupils eligible for free school meals (FSM) and pupil ethnicity), using data provided by schools. The data for English schools and Welsh schools will be processed separately in the SAIL databank and NPD secure environments.

We treat the primary outcome of attendance as confirmatory and all other analyses, including subgroup analyses, as exploratory.

For each subgroup analysis, we will add a dummy covariate to the primary specification of Equation 1, including an interaction term to consider the impact of group membership on the primary outcome (see Equation 2). We will be applying the BH procedure to determine which of the subgroups have a significant effect and calculating an effect size for those that are found to be significant.

Equation 2. Subgroup model

$$y_{ij} = \beta_0 + \beta_1 PreTest_{ij} + \beta_2 Treat_j + \beta_3 X_i + \beta_4 Treat_j \times X_i + \mu_j + e_{ij}$$

where

y_{ij} is the post-intervention attendance of *i*th pupil at school *j*

β_0 is the grand intercept

$PreTest_{1ij}$ is the pre-intervention attendance of *i*th pupil at school *j* (with associated regression coefficient β_1)

$Treat_j$ is a dummy variable = 1 if school *j* is intervention school (with associated regression coefficient β_2)

X_i is a dummy variable for subgroup analysis for pupil *i* (with associated regression coefficient β_3)

β_4 is the interaction regression coefficient

μ_j is the school-effect and it is assumed that and it is assumed that μ_j are normally distributed with mean 0 and variance σ_μ^2

e_{ij} is the residual for *i*th pupil at school *j* (and it is assumed the e_{ij} are normally distributed with mean 0 and variance σ_e^2)

Further analyses

Sensitivity analysis. We will perform sensitivity analysis to consider whether we obtain the same result for the impact of Grassroots on different kinds of absences. To do this, we will use the variable of school sessions missed due to illness from the NPD (England) and SAIL databank (Wales), applying the same model as Equation 1.

Mediator analysis. In order to further probe the theory of change, if the evaluation shows that Grassroots has an impact on attendance, we will conduct mediator analysis to analyse whether a reduction in bullying and peer-to-peer conflict is the mechanism by which this has

taken place. Depending on the results of the disciplinary report analysis, we will use the disciplinary report measure or the descriptive norms variable as a measure of peer-to-peer conflict.

Interim analyses and stopping rules

There are no interim analyses. The intervention is over a relatively short time scale and we will not have statistical data during this time that would provide us reason to stop the intervention. Stopping decisions will be judgements made as a result of ethical considerations that arise during the process and implementation evaluation.

Longitudinal follow-up analyses

The model described in Equation 1 will be used again for the follow-up analysis conducted one year following the intervention. No further data will have been collected regarding school compliance, and the same compliance analysis will be performed as applied to the primary outcome analysis. We will however survey schools regarding any attendance or antibullying interventions that they have carried out in the interim period and will conduct an additional analysis of the impact of this by adding a variable of further interventions to the primary outcome model. The longitudinal follow up will use attendance data for the Autumn term of the 2025-26 school year, collected from the National Pupil Database (England) and SAIL databank (Wales). The same measure of percentage school sessions attended in a term will be used against the same baseline data from the Autumn term of the 2023-24 school year.

Imbalance at baseline

We will report characteristics of the intervention and control groups for the following characteristics:

1. Overall school attendance (2022-23)
2. Persistent absence (2022-23)
3. Baseline measure of disciplinary reports (2023 data)
4. Baseline survey measures (SDQ, BCS-A, PCS, and feelings of safety in school, Oct. 2023)
5. Free school meal eligibility (2022-23)
6. Progress 8 score (2022-23)
7. Size of school (school roll 2022-23)
8. School admissions (selective or not)
9. School type (academy or not)

The characteristics will be reported for the baseline intervention and control groups at the point of randomisation. If schools drop out of the evaluation, these characteristics will also be reported for the control and intervention groups as analysed.

Characteristics 1, 3, 4, 5 have been selected because they are theorised to predict the impact of Grassroots and are already accounted for in the statistical analysis. In addition, we will report persistent absence (characteristic 2), because this is potentially an indicator of pupils who are more likely to benefit from the Grassroots intervention. We will report a further set of characteristics (6, 7, 8 and 9) to provide a fuller description of the schools that are participating in the Grassroots trial. Each of these characteristics may have an impact on the success of Grassroots.

For continuous data (1 to 6 above), we will report means and standard deviations. For categorical data (7 and 8 above), we will report the number and percentage of schools in each category. We will not report *p*-values for the differences, following CONSORT-10 guidelines. The reporting will show if randomisation has resulted in chance imbalances between the control and intervention groups and we follow current guidance in not adjusting our statistical models for this (de Boer et al., 2015).

The baseline characteristic data will be collated once the baseline surveys are completed, except for the disciplinary report data, which will be collected from schools in the summer of 2024.

Missing data

Our general approach to missing data is to analyse the extent of missing data and patterns of missing data, and to used observed data or multiple imputation (MI), conducting additional sensitivity analyses where necessary, depending on whether the missing data is judged to be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MRAM).

We will report the number of complete cases for each primary and secondary outcome, at both school-level and pupil-level. To investigate the patterns of missing data, we will use a multi-level logistic regression model. This will explore whether missingness for the primary outcome (a binary variable) can be predicted by the variables of our primary and secondary models (including the variables of our subgroup analysis of FSM eligibility and ethnicity), as well as feelings of safety in school (from our baseline survey) and gender (from data provided by school).

For our primary outcome (attendance), we have received consent from all schools in the trial to use attendance data from national databases and expect to have data for every school in the trial. At a pupil level, we expect a low level (<5%) of missing data, mainly due to pupils leaving a trial school during the school year. Pupils who join a trial school from a non-trial school during the school year will not be considered part of the trial. If the logistic regression model reveals no patterns of missingness in this data, we will conduct the primary outcome analysis with complete cases only. If we have reason to suspect that a few cases of missing

data may have a substantial influence on the results, we will apply multiple imputation (MI) to account for the missing data.

For our secondary outcomes, it is plausible that we will have missing data at a school level (if a school drops out of the trial or is unable to complete one of the surveys). We also expect to have a higher amount of missing data due to pupil absence when surveys are conducted. This missingness is unlikely to be MCAR because students who are experiencing bullying and peer conflict are less likely to be at school to complete both pre-intervention and post-intervention surveys. Because our secondary outcomes are exploratory, we will not conduct logistic regression for these, but concentrate our efforts on exploring the patterns and reasons for missingness for the primary outcome.

Compliance

Compliance with the intervention will be defined at the school level. A school will be considered to be compliant if:

- There has been a delivery of a coherent Grassroots programme within the school, with a minimum of 5 sessions, including Grassroots Day
- The attendance at Changemakers sessions have included a minimum of 40% of seed group pupils overall and at least 33% of seed group pupils from each year group in KS3 (Year 7, 8 and 9).

These criteria have been formed in agreement with the developer during the pilot stage, reflecting the experience of the intervention in the US and therefore realistic expectations for the delivery team (Paluk et al, 2016)

The primary analysis will take an intention-to-treat approach and compare outcomes for schools in the control and intervention groups regardless of compliance. A further analysis using an Instrumental Variables (IV) approach will be conducted to see if outcomes differ within the intervention group according to compliance (Angrist & Imbens, 1995). This will use a Two Stage Least Square (2SLS) approach. We will estimate a (first stage) model of compliance as follows:

$$Comply_j = \beta_0 + \beta_1 Treat_j + \beta_2 PreTest_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

Where $Comply_j$ is the binary compliance variable defined above. The predicted values of $Comply_j$ from the first stage are used in the estimation of the second stage model of our outcome measure y_{ij} which is specified as follows:

$$y_{ij} = \beta_0 + \beta_1 \widehat{Comply}_j + \beta_2 PreTest_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

where $\widehat{Comply}_j$ are the predicted values derived from the first stage model. Our primary outcome of interest will be β_1, which estimates the effect of the intervention among compliers. In line with YEF's analysis guidance (2021), results for the first stage will be reported alongside the correlation between the outcomes and compliance and an F-test.

Intracluster correlations (ICCs)

We will measure how much of the total variation in the primary outcome (attendance) is attributable to variation between schools. The standard ICC we will use to do this is described as the 'adjusted ICC' by Nakagawa *et al.* (2017), 'ICC1' by Bliese (1998) and 'ICC(1,1)' by Shrout and Fleiss (1979), and is calculated as:

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

Where σ_u^2 and σ_e^2 are the variances in the school effect and the individual pupil effect respectively. These will be calculated using the empty model given by Equation 3.

Equation 3. The empty model

$$y_{ij} = \beta_0 + \beta_1 x_j + \mu_j + e_{ij}$$

where

y_{ij} is the increase in attendance of *i*th pupil at school *j*

β_0 is the grand intercept

x_j = 1 if school *j* is an intervention school (with associated regression coefficient β_1).

μ_j is the school effect and it is assumed that $\mu_j \sim N(0, \sigma_u^2)$

e_{ij} is the residual for *i*th pupil at school *j* and it is assumed that $e_{ij} \sim N(0, \sigma_e^2)$

Presentation of outcomes

Effect sizes will be calculated using the Hedge's *g* for as per the YEF statistical analysis guidance for evaluations (YEF, 2021):

$$ES = \frac{(\bar{Y}_t - \bar{Y}_c)_{adj}}{s^*}$$

Where $(\bar{Y}_t - \bar{Y}_c)_{adj}$ denotes the difference in the means between the trial and intervention groups adjusting for the baseline predictor (recovered from the β_2 coefficient in Equation 1 of the primary model) and s^* denotes the pooled unconditional variance of the two groups:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Where the control group has sample size n_1 and standard deviation s_1, and the intervention group has sample size n_2 and standard deviation s_2. We will calculate ninety-five per cent confidence intervals (95% CIs) of the effect size by inputting the upper and lower confidence limits of β_2 from the primary model into the effect size formula.

Effect sizes and their associated CIs will be calculated for all outcomes that have been found to be significant using the BH procedure. We will provide conversions from effect sizes to number of school sessions missed to convey the impact meaningfully.

References

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

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., and Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. M*ethodol. Eur. J. Res. Methods Behav. Soc. Sci*. 10, 1–11.

Bennett D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5), 464–469.

Bliese, P. D. (1998). Group Size, ICC Values, and Group-Level Correlations: A Simulation. Organizational Research Methods, 1(4), 355–373. https://doi.org/10.1177/109442819814001

de Boer, M.R., Waterlander, W.E., Kuijper, L.D. et al. (2015) Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act* 12, 4. https://doi.org/10.1186/s12966-015-0162-z.

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical methods in medical research,* 16(3), 259–275. https://doi.org/10.1177/0962280206075303

Demack, S. (2019). Does the classroom level matter in the design of educational trials? A theoretical & empirical review. EEF Research Paper No. 003.

Epstein, S., Roberts, E., Sedgwick, R., Polling, C., Finning, K., Ford, T., Dutta, R., & Downs, J. (2020). School absenteeism as a risk factor for self-harm and suicidal ideation in children and adolescents: a systematic review and meta-analysis. *European child & adolescent psychiatry*, 29(9), 1175–1194. https://doi.org/10.1007/s00787-019-01327-3

Francis-Devine, B., Danechi, S., Malik, S. (2023). Food poverty: Households, food banks and free school meals. Research briefing. House of Commons Library. https://researchbriefings.files.parliament.uk/documents/CBP-9209/CBP-9209.pdf

Gamble, C. & Hollis, S. (2005). Uncertainty method improved on best–worst case analysis in a binary meta-analysis. *Journal of Clinical Epidemiology,* 58(6), 570–588. https://doi.org/10.1016/j.jclinepi.2004.09.013

Hutchison, D. & Styles, B. (2010). A Guide to Running Randomised Controlled Trials for Educational Researchers. Slough: NFER.

Kreft, I., and de Leeuw, J. (1998). Introducing Multilevel Modeling. London: Sage Publications.

Marsee, M.A., Frick, P.J., Barry, C.T., Kimonis, E.R., Munoz-Centifanti, L.C., & Aucoin, K.J. (2014). Profiles of the forms and functions of self-reported aggression in three adolescent samples. *Development and Psychopathology*, 26, 705-720.

Mavridis, D. & White, I. R. (2019). Dealing with missing outcome data in meta-analysis. *Research Synthesis Methods* 11(1) 2-12. https://doi.org/10.1002/jrsm.1349

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface,* 14(134), 20170213. doi:10.1098/rsif.2017.0213

Özbey, H & Öznur Başdaş, Ö. (2020). Psychometric properties of the Turkish version of the Bullying and Cyber Bullying Scale for Adolescents (BCS-A), *Psychiatry Research*, 289: 112994, https://doi.org/10.1016/j.psychres.2020.112994

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. Proceedings of the National Academy of Sciences, 113(3), 566–571. https://doi.org/10.1073/pnas.1514483113

Pechorro, P., Hayes, N., Gonçalves, M. (2012). Portuguese Validation of the Brief Peer Conflict Scale: Validity, Reliability and Invariance across Gender. *J Psychopathol Behav Assess*, 43, 216–226. https://doi.org/10.1007/s10862-020-09826-3

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological bulletin, 86(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Singh, A., Uwimpuhwe, G, Vallis, D., Akhter, N., Coolen-Maturi, T., Einbeck, J., Higgins, S., Culliney, M., Demack, S. (2023). Improving power calculations in educational trials. Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Work_Package_2023-WP6_18_09_2023_FINAL.pdf?v=1714993988Thomas, H.J., Scott, J.G., Coates, J.M. and Connor, J.P. (2019), Development and validation of the Bullying and Cyberbullying Scale for Adolescents: A multi-dimensional measurement model. *Br J Educ Psychol,* 89: 75-94. https://doi.org/10.1111/bjep.12223

Timpson, E. (2019). Timpson Review of School Exclusion. London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/807862/Timpson_review.pdf

Van Breukelen G. J. (2006). ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies [corrected]. Journal of clinical epidemiology, 59(9), 920–925. https://doi.org/10.1016/j.jclinepi.2006.02.007

YEF (2021). *Analysis guidance*. Youth Endowment Fund. Last updated June 2021. https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf

Appendix

This appendix contains indicative syntax for the implementation of the approach outlined in the statistical analysis plan, including the code already used in the randomisation procedure. A full account of all syntax used in the analysis will be appended to the final trial report.

Randomisation

```
#set the random seed for replicability
set.seed(1)

#generate random numbers for each school
school_list <- runif(120)
```

Model building

```
#primary outcome model
lme.primary <- lmer(primary_outcome ~ 1 + allocation + pre_attend +
(1|SchoolID), data = attend_data)
```

Subgroup modelling

```
#subgroup analysis model
lme.subgroup <- lmer(primary_outcome ~ 1 + allocation + pre_attend +
subgroup + subgroup:allocation + (1|SchoolID), data = attend_data)
```

Intraclass correlation (ICC) estimation

```
#null model for ICC
null_model <- lmer(primary_outcome ~ 1 + (1|SchoolID), data =
attend_data)

#calculate the adjusted ICC
performance::icc(null_model)
```

Missing data

```
# create indicator of missing data
attend_data$outcome.miss <- 0
attend_data$outcome.miss[is.na(attend_data$primary_outcome) == TRUE]
<-1

# logistic regression to see if missingness can be predicted
miss.glmer <- glmer(outcome.miss ~ allocation + pre_attend +
FSM_eligibility + ethnicity + feelsafe + missed_sch + gender +
(1|SchoolID), family = binomial(logit), data = attend_data)
```

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413