Face It: Pilot study

Statistical Analysis Plan

18/12/2023

Lilli Wagstaff,
Alice Worsley,
Jack Martin,
Tom McBride.

Acknowledgements

We are grateful to colleagues from Khulisa (including Iman Haji and Derellanne Knowles) for their assistance with and input into the early stages of this evaluation. We are also grateful to colleagues from the Behavioural Insights Team (particularly Dr Patrick Taylor, and Leonie Nicks), the Youth Endowment Fund (particularly Dr Mollie Bourne, Dr Daniel Acquah, and Dr Amy Wells) and Dr Nick Axford (University of Plymouth) for their helpful advice and feedback.

We are grateful to the Youth Endowment Fund (originally funded by an endowment from the Home Office) and Stuart Roden for the funding which supports this work, and to the Behavioural Insights Team and Nesta for their support and role in administering this funding.

THE
BEHAVIOURAL
INSIGHTS
TEAM

YOUTH
ENDOWMENT
FUND

nesta

About

This report was first published in December 2023, and is available to download as a free PDF at: https://www.bi.team/wp-content/uploads/2023/12/Face-It-SAP.pdf

This work is being undertaken by the Ending Youth Violence Lab at the Behavioural Insights Team. The Behavioural Insights Team (BIT) is a global social purpose company that generates and applies behavioural insights to inform policy, improve public services and deliver results for citizens and society. With company number 08567792.

Contents

Project title	Face It Evaluation - Pilot study
Developer (Institution)	Khulisa
Evaluator (Institution)	Ending Youth Violence Lab
Principal investigator(s)	Tom McBride
SAP author(s)	Lilli Wagstaff, Alice Worsley, Jack Martin, Tom McBride
Trial design	Two-armed individually randomised controlled trial
Trial type	Pilot
Evaluation setting	5 schools in London
Target group	13-15 year-olds living in London at risk of school exclusion
Planned number of participants	5 schools, approximately 160 young people
Primary outcome and data source	Behavioural difficulties: ● Self-reported Strengths and Difficulties Questionnaire: Total difficulties score (SDQ) (at post-programme)
Secondary outcome and data source	Behavioural difficulties: ● Self-reported SDQ: Total difficulties score (at 3-month follow up) ● Parent-reported SDQ: Total difficulties score (post-intervention; 3-month follow-up) ● Self-reported SDQ: Externalising and internalising scores (post-intervention; 3-month follow-up) ● Parent-reported SDQ: Externalising and internalising scores (post-intervention; 3-month follow-up) Offending: ● The Self-Report Delinquency Scale (post-intervention; 3-month follow-up) Victimisation: ● Problem Behaviour Frequency Scale (post-intervention; 3-month follow-up) Resilience ● Children’s Hope Scale (post-intervention; 3-month follow-up)

	Emotional regulation ● The Emotional Regulation Questionnaire (post-intervention; 3-month follow-up) Social and emotional wellbeing ● Short Warwick Edinburgh Mental Wellbeing Scale (SWEMWBS) (post-intervention; 3-month follow-up) School attendance ● ImpactEd data (3-month follow-up) School exclusions ● ImpactEd data (3-month follow-up)

1. Introduction

Face It is a schools-based intervention, designed and implemented by Khulisa. It is an intensive therapeutic group programme for young people, focused on developing social and emotional skills and designed to explore the root causes of emotional distress. It combines creative techniques like storytelling, art, debating, and role-play, with the latest developments in neuroscience. Face It has indicated early promise as an intervention to improve social and emotional skills. The intervention has demonstrated that it can recruit and retain participants, and qualitative work indicates that the programme is well-regarded by participants.

While Khulisa is committed to evaluation and evidence building, a randomised control trial to establish impact has not yet been conducted. Before any future full-scale randomised control trial (RCT), it is important that a small-scale pilot trial is conducted in advance to support and inform this work - to test and improve evaluation procedures such as randomisation and data collection, and to generate useful information around sample size determination.

Subsequently, the Lab will be conducting a small-scale individual-level pilot RCT, where eligible pupils are randomly assigned to one of two arms: the intervention arm where pupils receive Face It, and the control arm where pupils receive business-as-usual (BAU). The primary objective of the study is to understand the feasibility of running a full scale efficacy trial to evaluate the impact of Face It. We will look at questions like the feasibility of randomisation and data collection, as well as the acceptability of an evaluation in this context. For more detailed information about the intervention and the trial, please see the evaluation protocol. This statistical analysis plan develops this protocol by providing more detail on quantitative data collection, outcome analysis, subgroup analysis, and treatment of missing data.

2. Design overview

Trial design, including number of arms		Two-arm, individually randomised
Unit of randomisation		Individual participant: pupil
Stratification variables (if applicable)		Externalising vs internalising behavioural problems Gender
Primary outcome	variable	*Behavioural difficulties (total)*
	measure (instrument, scale, source)	*SDQ: total difficulties, 0-40, post-programme self-assessment*
Secondary outcome(s)	variable(s)	*Behavioural difficulties (total)* *Behavioural difficulties (internalising and externalising)* *Offending* *Victimisation* *Resilience* *Emotional regulation* *School attendance* *School exclusions*
	measure(s) (instrument, scale, source)	*SDQ: total difficulties, 0-40, 3-month follow up self-assessment* *SDQ: internalising score, 0-20, post-programme and 3-month follow-up self-assessment* *SDQ: externalising score, 0-20, post-programme and 3-month follow-up self-assessment* *SDQ: total difficulties, 0-40, post-programme and 3-month follow up parent-assessment* *SDQ: internalising score, 0-20, post-programme and 3-month follow-up parent-assessment*

		SDQ: externalising score, 0-20, post-programme and 3-month follow-up parent-assessment *The Self-Report Delinquency Scale, 0-19, post-programme and 3-month follow up self-assessment* *Problem Behaviour Frequency Scale, 6-36, post-programme and 3-month follow up self-assessment* *Children's Hope Scale, 6-36, post-programme and 3-month follow up self-assessment* *The Emotional Regulation Questionnaire, 0-10, post-programme and 3-month follow up self-assessment* *Social and emotional wellbeing, SWEMWBS, 7-35, post-programme and 3-month follow up self-assessment* *Number of sessions attended/total number of sessions during the pilot period, ImpactEd data* *Number of exclusions during the pilot period, ImpactEd data*
Baseline for primary outcome	**variable**	*Behavioural difficulties*
	measure (instrument, scale, source)	*SDQ: total difficulties, 0-40, pre-programme self-assessment*
Baseline for secondary outcome	**For all the above**	*Pre-programme self-assessments/baseline data collections*

3. Sample size calculations overview

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		0.45	
Pre-test/ post-test correlations	level 1 (participant)	N/A	
	level 2 (cluster)	N/A	
Intracluster correlations (ICCs)	level 1 (participant)	N/A	
	level 3 (cluster)	N/A	
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		two-sided	
Average cluster size		N/A	
Number of clusters	intervention	N/A	
	control	N/A	
	total	N/A	
Number of participants	intervention	80	
	control	80	
	total	160	

As this is a pilot trial and its primary objective is to investigate evaluability and to test evaluation processes (rather than to identify impact), we did not identify what we believed to be a likely effect size for the Face It programme and then seek to recruit a sample large enough to identify that effect.

Instead, the sample size was agreed with Khulisa based on their experience of recruitment, their capacity to deliver the intervention, and what we believed to be a proportionate sample size for a pilot trial. Through Khulisa's experience of delivering Face It, they have found that the optimal group size for delivery is 8-10 pupils. Additionally, it was decided that recruiting more than 5 schools in the pilot timelines would be unfeasible. However, to increase sample size, it was agreed that we would deliver the programme to two cohorts within each school. Therefore, we anticipate that it is feasible to deliver to 80-100 pupils total.

Based on this sample size, we conducted power calculations to determine what MDES we may observe. We used means and standard deviations for our primary outcome measure (self-reported SDQ: total difficulties), from previous trials to estimate our MDES in STATA. The results showed an MDES of 0.45.

Our sample will consist of pupils who are deemed eligible by Khulisa. Schools will refer pupils at risk of offending, exploitation and school exclusion to Khulisa. They will use the referral form and the participant profile to identify pupils[1]. Once Khulisa have received the referrals, they will review them against the eligibility criteria set out.

If a school refers more than 20 pupils, we will conduct a two-stage randomisation process where we first randomly select a sub-sample of 20 eligible pupils and then randomly allocate to treatment or control within that sub-sample. This is in order to promote optimal group sizes, and to prevent a situation where more young people are assigned to the intervention group than Khulisa have the capacity to deliver to.

[1] The participant profile is a document written by Khulisa which gives guidance to schools on how to successfully and safely recruit participants, and on what characteristics make young people eligible or ineligible for the programme.

4. Quantitative Research Activities

Acquisition of quantitative data during the pilot study will occur during the administration of **outcome surveys** to both pupils and their caregivers, and during collection of **programme administrative data** and **ImpactEd data**. A brief summary of the relevant data collected as part of each is included below. For further details see Section 5 of the study protocol.

Outcome surveys

Pupils who have been deemed eligible for Face It will be invited to complete an outcome survey at baseline, prior to randomisation. We will then invite all pupils who completed the baseline survey to complete a post-programme survey and a 3-month follow up survey.

We will also be sending an outcome survey to the parents of all eligible pupils, at baseline, post-programme and a 3-month follow up.

Given the primary objective of the pilot is to understand the feasibility of a full scale evaluation, we are primarily administering outcome surveys to obtain information about response and completion rates. These quantitative metrics will help us to identify whether particular measures, or questions, have lower rates of engagement of completion than others, and/or whether caregivers with certain demographic characteristics engage less with any aspect of the survey. Additionally, we will conduct effectiveness analysis to get a preliminary sense of the effectiveness of Face It, acknowledging that our study is unlikely to be powered to identify statistically significant effects, and may produce imprecise estimates of effect

ImpactEd data

We will use ImpactEd data to collect data on school attendance and exclusions at baseline and 3-month follow up.

Programme administrative data

We will analyse administrative data from the following sources:

1. Khulisa referral forms
2. Khulisa programme administration data (including attendance lists)
3. Khulisa fidelity monitoring data[2]

[2] Facilitators are required to complete a 5-day session guide provided by Khulisa.

5. Analysis

Feasibility, acceptability and evaluability analysis

The primary objective of the pilot trial is to establish the feasibility, acceptability, and evaluability of evaluating Face It. To answer these questions, we will conduct the following analyses.

Outcome surveys

We will report overall survey response, attrition, and completion rates, as well as attrition and completion rates for each specific outcome measure and comparisons across treatment and control, based on data collected from both the children/young people (CYP) and the caregiver outcome surveys. Where appropriate these will be broken down by pre-test, post-test, and follow-up data collection points. Specifically, we will report:

- **Response rate** - this will involve calculating, as a percentage, the number of participants who started the survey, compared to the total number of participants who were invited to complete the survey.
  - $Survey\ response\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ started\ survey}{\#\ participants\ who\ were\ invited\ to\ complete\ survey}\right)\ *\ 100$
- **Attrition rate** - this will involve calculating, as a percentage, the number of participants who finished the survey, compared to the total number of participants who were randomised and entered the study (and broken down by treatment and control groups).
  - $Attrition\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ finished\ survey}{\#\ participants\ who\ were\ randomised\ to\ study\ groups}\right)\ *\ 100$
- **Attrition rates for specific measures** will be calculated as follows:
  - $Measure\ attrition\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ completed\ measure}{\#\ participants\ who\ were\ randomised\ to\ study\ groups}\right)\ *\ 100$
- **Completion rate** - this will involve calculating, as a percentage, the number of participants who finished the survey, compared to the total number of participants who started the survey (and broken down by treatment and control groups).
  - $Survey\ completion\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ finished\ survey}{\#\ participants\ who\ started\ survey}\right)\ *\ 100$
- **Completion rates for specific measures** will be calculated as follows:
  - $Measure\ completion\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ completed\ measure}{\#\ participants\ who\ finished\ survey}\right)\ *\ 100$
- **Completion rates for specific questions** will be calculated as follows:
  - $Question\ completion\ rate\ (\%)\ =\ \left(\frac{\#\ participants\ who\ answered\ question}{\#\ participants\ who\ finished\ survey}\right)\ *\ 100$

In addition, alongside the outcome surveys, we explore:

- **CYP perception of the programme content and delivery -** this will involve calculating the mean and standard deviation of scores, based on a short feedback survey administered alongside the post-programme outcome survey.
    - $Satisfaction\ with\ practitioners\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$ [3]
    - $Satisfaction\ with\ duration\ of\ programme\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$
    - $Satisfaction\ with\ group\ format\ delivery\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$
    - $Satisfaction\ with\ taught\ content\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$
    - $Satisfaction\ with\ programme\ activities\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$
    - $Perception\ of\ programme\ meeting\ CYP\ needs\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$
    - $Overall\ satisfaction\ with\ programme\ (mean\ score\ between\ [1, 5])\ =\ Adjusted\ mean\ score$

- **Alternative provision -** we will report counts and proportions of CYP who have been involved in other programmes during the trial period, based on a multiple choice question listing other programmes offered in the school. We will break this down by treatment and control group.

Administrative data analysis

Analysis of the programme administrative data will include the calculation of descriptive statistics on key variables including:

- **Proportion of schools recruited to receive Face It -** this will involve calculating, as a percentage, the number of recruited schools, compared to the total number of schools approached by Khulisa.
    - $Schools\ recruited\ (\%)\ =\ (\frac{\#\ schools\ successfully\ recruited}{\#\ schools\ approached})\ *\ 100$

- **Retention / drop-out rates of schools in programme & evaluation -** this will involve calculating, as a percentage, the number of schools where the programme is delivered in full and where there is full participation in the evaluation (i.e. post-intervention data collection sessions occur), compared to the total number of successfully recruited schools.
    - $School\ retention\ (\%)\ =\ (\frac{\#\ schools\ w/\ completed\ delivery\ and\ post-test\ data}{\#\ schools\ successfully\ recruited})\ *\ 100$

- **Number of referrals received by Khulisa for CYP to receive Face It -** this will involve identifying the total count of referrals received in each school.

- **Proportion of referred CYP deemed eligible for Face It -** this will involve calculating, as a percentage, the number of CYP deemed eligible for Face It, compared to to the total number of referred CYP.
    - $CYP\ eligible\ (\%)\ =\ (\frac{\#\ eligible\ referred\ CYP}{\#\ total\ CYP\ referred\ to\ Face\ It})\ *\ 100$

[3] The **adjusted mean score** will be calculated using simple ordinal encoding of the 5-point scale answers in the feedback survey. The answer options - a Likert scale ranging from very unhelpful/very happy to very helpful/very happy - will be encoded as 1, 2, 3, 4, and 5 respectively.

- **Proportion of eligible CYP who are offered Face It who agree to the invention and evaluation** - this will involve calculating, as a percentage, the number of CYP who consent to take part in the pilot study, compared to to the total number of CYP who are considered eligible to receive and are offered the programme.
    - $CYP\ takeup\ (\%)\ =\ (\frac{\#\ CYP\ who\ consent\ to\ receiving\ Face\ It}{\#\ eligible\ CYP\ offered\ \ Face\ It})\ *\ 100$
- **Retention / drop-out rates of CYP receiving Face It -** this will involve calculating, as a percentage, the number of CYP who complete the programme, compared to the total number of CYP who consent to receiving Face It.[4]
    - $CYP\ \ retention\ (\%)\ =\ (\frac{\#\ CYP\ who\ complete\ programme}{\#\ \ CYP\ who\ consent\ to\ receiving\ Face\ It})\ *\ 100$
- **Attendance rate at Face It sessions -** this will involve calculating, as a percentage, the number of CYP who attended each Face It session[5], compared to the number of CYP who were expected (i.e. based on CYP take-up/enrollment) to attend these sessions.
    - $Session\ CYP\ attendance\ (\%)\ =\ (\frac{\#\ CYP\ who\ attended\ the\ session}{\#\ \ CYP\ expected\ to\ attend\ the\ session})\ *\ 100$
- **Programme dosage/fidelity data -** this will involve calculating the mean percentage of actual content delivered in Face It sessions to a cohort of CYP, compared to the amount of planned content to be delivered, as captured and recorded by the 5-day session plan.
- **Compliance with randomisation** - this will involve calculating, as a percentage, the number of control participants who attend a Face It session, using the attendance sheets.
    - $Randomisation\ compliance\ (\%)\ =\ (\frac{\#\ control\ CYP\ who\ attended\ a\ \ session}{\#\ \ CYP\ in\ the\ control\ group})\ *\ 100$
- **Number of complaints -** this will involve identifying the total count of complaints received in each school.

For all metrics above, we will use descriptive statistics to disaggregate these metrics and explore whether they vary by key participant characteristics, such as:

- School type/location
- Child/young person gender
- Caregiver gender
- Child/young person ethnicity
- Caregiver ethnicity

[4] Khulisa defines ‘completing’ the programme as attending all 5 days of the 5-day intensive programme. This means that young people are able to miss pre-programme and post-programme 1:1 and/or group sessions and would still be considered to complete the programme if they attend all of the 5-day programme.

[5] Note that we define a ‘session’ as a 1:1 or a day of the 5-day programme. For example, day 1 of 5 will count as one session.

- Family socioeconomic status
- Child/young person age

Effectiveness analysis

In addition to the analysis of the feasibility, acceptability and evaluability of Face It, we will conduct impact analysis of outcome data, in order to further inform monitoring criteria and our resulting recommendation to YEF.

Our analysis will help us understand if there is sufficient evidence of impact to justify a larger and more robust efficacy trial. Because this is a pilot study with a small sample size, we will have to interpret any statistical results with caution.

All outcome data will be analysed using an intention to treat (ITT) analysis using Stata or RStudio.

Given the high number of outcomes and comparisons we will be making, we will be adjusting for multiple comparisons. While this is not strictly necessary in a pilot study where the aim of the trial is not to estimate impact, we feel it is important to replicate the analytical approach which would be used in a full scale efficacy trial. We will apply the Benjamini-Hochberg method to correct the p-values reported for all outcomes.

Primary outcome analysis

Our primary outcome is the post-programme self-reported SDQ (total difficulties).

Analysis will be carried out using an ordinary least squares (OLS) regression, detailed below:

$$(1)\ Y_i = \beta_0 + \beta_1 T_i + \beta_2 PreSDQ_i + \beta_3 X_i + \epsilon_i$$

Where:

- Y_i is the post-programme SDQ score for individual i;
- T_i is a binary indicator for the treatment for individual i (1 if the pupil is in the intervention arm and 0 if not);
- $PreSDQ_i$ is the baseline SDQ score for individual i;
- X_i is a vector of pupil covariates including gender, ethnicity, FSM status, and allocation reason[6]; and
- ε_i is the robust error term.

[6] Programme groups will include young people displaying externalising and internalising behaviours with an approximate ratio of 80:20, as Khulisa views this balance as an essential component of building a group dynamic. 'Allocation reason' refers to the reason a child/young person was referred to the programme - i.e. whether they have been categorised as having internalising or externalising issues.

Secondary outcome analysis

We will investigate a number of secondary outcomes, listed below. Data on these will be collected at three time points: baseline, post-programme and 3-month follow up.

Below we detail the regressions we will use for the secondary outcome analyses, noting they use the same specification as the primary analysis.

Post-programme secondary outcome analysis

All variables below will be analysed using the (2) regression specification.

- The Self-Report Delinquency Scale
- Problem Behaviour Frequency Scale
- SDQ (parent assessment)
- The Children’s Hope Scale
- The Emotional Regulation Questionnaire
- SWEMWBS
- School Exclusions
- School Attendance

$$(2)\ Y_i = \beta_0 + \beta_1 T_i + \beta_2 Pre_i + \beta_3 X_i + \epsilon_i$$

Where:

- Y_i is the post-programme score or value for individual i;
- T_i is a binary indicator for the treatment for individual i (1 if the pupil is in the intervention arm and 0 if not);
- Pre_i is the baseline score or value for individual i;
- X_i is a vector of pupil covariates including gender, ethnicity, FSM status, and allocation reason[7]; and
- ε_i is the robust error term.

3-month follow-up secondary outcome analysis

All variables below will be analysed using the (3) regression specification.

- SDQ (self assessment)
- The Self-Report Delinquency Scale
- Problem Behaviour Frequency Scale
- SDQ (parent assessment)
- The Children’s Hope Scale
- The Emotional Regulation Questionnaire
- SWEMWBS
- School Exclusions

[7] As above.

- School Attendance

$$(3)\ Y_i = \beta_0 + \beta_1 T_i + \beta_2 Pre_i + \beta_3 X_i + \epsilon_i$$

Where:

- Y_i is the 3-month follow-up score or value for individual i;
- T_i is a binary indicator for the treatment for individual i (1 if the pupil is in the intervention arm and 0 if not);
- Pre_i is the baseline score or value for individual i;
- X_i is a vector of pupil covariates including gender, ethnicity, FSM status, and allocation reason[8]; and
- ε_i is the robust error term

Subgroup analyses

We will conduct two subgroup analyses for the 80% of young people showing externalising behaviour to test whether these outcomes vary from the 20% displaying internalising behaviour.

1. Estimate the model specified in the primary analysis on the subsample of externalising pupils;
2. Estimate a similar model including an interaction term for allocation reason, using the entire sample (see Equation 4 below).

Both approaches will estimate the effect size for externalising pupils, but the latter uses information from the whole sample. Under ideal conditions, the total treatment effect in both should be analogous. The results for both will be compared and reported in the Appendix as a robustness check for this subgroup analysis. We will explore any difference and discuss its implications for the results.

$$(4)\ Y_i = \beta_0 + \beta_1 T_i + \beta_2 Externalising_i + \beta_3 (Externalising_i * T_i) + \beta_5 X_i + \epsilon_i$$

Where:

- Y_i is the post-programme SDQ score for individual i;
- T_i is a binary indicator for the treatment for individual i (1 if the pupil is in the intervention arm and 0 if not);
- $Externalising_i$ is a binary indicator equal to 1 if the pupil was referred due to their externalising behaviour and 0 if not);

[8] 'Allocation reason' refers to the reason a child/young person was referred to the programme - i.e. whether they have been categorised as having internalising or externalising issues.

- X_i is a vector of pupil covariates including gender, ethnicity, FSM status, and allocation reason[9]; and
- ε_i is the robust error term.

We will report and compare the effect sizes of both models, the restricted sample and the interaction term. The effect sizes will be reported in terms of Hedges' G.

In terms of how young people are placed into either subgroup, teachers are asked to select any of the following behavioural indicators that apply to a given pupil in their referral form:

- Externalised behavioural indicators:
  - Disruptive or antisocial behaviour
  - Unpredictable outbursts of anger
  - Verbally abusive
  - Threatening behaviour

- Internalised behaviour indicators
  - Isolated in lessons or from peer group
  - Anxious and withdrawn
  - Historic experience of self-harm (over 6 months)
  - Historic eating disorder (over 1 year)

A count is created where selection of any externalised behaviour indicator is double weighted and selection of any internalised behaviour indicator is single weighted. If the weighted count is greater for externalised behavioural indicators, the pupil is defined as showing externalising behaviour, and if the weighted count is greater for internalised behavioural indicators, the pupil is defined as showing internalising behaviour.

Further analyses

Robustness check

We will conduct a robustness check where we remove participants who completed the survey in less than 2 minutes. This is because we anticipate that this means they selected responses at random. The specification will follow the same as that of the primary analysis. The results for both sets of analyses (with, and without those completing in under 2 minutes) will be compared and reported in the Appendix as a robustness check. We will explore any difference in findings and discuss its implications for the results

[9] 'Allocation reason' refers to the reason a child/young person was referred to the programme - i.e. whether they have been categorised as having internalising or externalising issues.

Interim analyses and stopping rules

Given that this is a pilot study, we do not have any interim analyses or stopping rules. Instead, we have a number of monitoring and success criteria that we will continue to check throughout the pilot. For more details, see section 4 in the study protocol.

Imbalance at baseline

We will present a table showing baseline test scores and demographic covariates (gender, ethnicity, FSM status, and allocation reason) for both the intervention and control groups. We will test for balance by comparing normalised differences in means for continuous variables and counts/percentages for categorical variables. We will produce these statistics at both randomisation and in the analysis.

The normalised difference is defined as the difference in means between the two groups, divided by the pooled standard deviation. Normalised differences with a magnitude of 0.1 or less indicate a negligible correlation between the covariate and assignment to treatment group, which can usually be addressed through covariate adjustment in the regression (Austin, 2009). We will report on any differences and discuss the implications for the primary and secondary analyses.

Missing data

The two main types of missing data are:

A. Missing pre-treatment covariates
B. Missing outcome data

Below, we detail our approach to handle both types. For both, we will report the number of missing observations and will try to establish the mechanism behind missingness for each variable. Data can be missing completely at random (MCAR), missing at random (MAR), or missing not a random (MNAR), and the approach to analysis will vary depending on the type.

A - Missing pre-treatment covariates

Pre-treatment covariates will be coming from two sources: *baseline survey data* and the *referral data*. We anticipate that we will have limited missing data from either source.

The *baseline survey data* will be collected pre-randomisation, making it necessary for participants to have submitted a baseline survey to be included in the sample.

Complete *referral data*, on the other hand, is a requirement for being eligible for the trial. Khulisa will need complete data to be able to assess eligibility for the programme.

However, in the unlikely event that more than 5% of pre-treatment covariates are missing, we will try to establish which variables are predictive of the missing data. To do this, we will create a new variable that is a binary indicator of missingness and look for its predictors using a logistic regression model to establish correlations with the other variables in the dataset. For all covariates including gender, ethnicity, FSM status, and allocation reason, missing data will be modelled as follows:

$$M_i \sim \; binomial(p_i); \; logit(p_i) \; = \beta_0 + \beta_1 X_i$$

where:

- M_i is the binary variable for missingness (equal to 1 if missing and 0 if not missing);
- p_i is the probability that a given observation is missing the covariate in question
- X_i is a vector of the remaining pupil covariates

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

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

Any imputation of covariates will be restricted to the primary analysis and will only be carried out when more than 5% of the data is missing. Schultz and Grimes (2002) suggest that when less than 5% of data is missing, there is likely to be little bias introduced to estimated treatment effects, so we have adopted this threshold here.[10]

B - Missing outcome data

Given we have outcome data from two main sources, *survey data* and *ImpactEd data*, we anticipate there being a number of reasons why we may have missing data. We highlight these below.

For both data sources, we will implore a similar approach as the one outlined in the previous section. The regression specification will be:

[10] This is also in line with the convention provided in EEF's guidance on statistical analysis.

$$M_i \sim \ binomial(p_i);\ logit(p_i) \ = \beta_0 + \beta_1 X_i$$

where:

- M_i is the binary variable for missingness (equal to 1 if missing and 0 if not missing);
- p_i is the probability that a given observation is missing the outcome in question
- X_i is a vector of the pupil covariates

P-values below 0.05 will be considered evidence of missingness being conditional on covariates or MAR (missing at random). In this case, complete case analysis may yield biased estimates so we will estimate a model including only those covariates and interpret the results.

If there is no evidence that data is missing conditional on observables, data missingness may be MCAR (missing completely at random) or MNAR (missing not at random).

Data is MCAR when missingness is uncorrelated with both observables and unobservables. This could occur in the case of pupils missing a test because they were sick, or because they left the school. Whether missing data is correlated (or not) with unobservables will depend on the context of the trial. Whenever possible, we will try to gather information from the schools on the reason for a missing test result during baseline and endline data collection and try to identify whether it was a case of persistent or a one-time absence, a withdrawal from the trial or the evaluation, or that the pupil left the school.

In the case of MCAR, we will run a complete case analysis as this will yield unbiased results.

If data is MNAR (missingness is correlated with unobservables), multiple imputation will not correct for the bias. In this case, we may choose to use sensitivity analysis to compare with the main estimates.

Missing survey data

Outcome data collected via surveys may be missing for the following reasons:

1. **Participants did not receive the follow-up surveys:** For the pupils, this will primarily be because they were sick or not in school on the day of data collection. It may also be because schools failed to schedule the sessions at a time that worked for all pupils. For caregivers, this will primarily be because the contact information we have for them is not accurate or valid.
2. **Participants receive the follow-up surveys but do not complete it:** For pupils, we anticipate that this will be rare, since the sessions are in school time and they will be required to sit through the session regardless of if they complete it or not. However, we anticipate that this will be the primary source of missing data for caregivers, since we are asking them to complete the survey on their own time, without anyone instructing them to begin. Additionally, they may not see the message with the link.

3. **Participants do not finish the follow-up surveys:** We anticipate that, especially for the caregivers, there may be a large proportion of participants who begin the survey but do not complete it.
4. **Participants skip certain questions within scales:** Since we have decided to not make answering all questions mandatory, we expect the main source of missing data to come from participants skipping certain questions. This may be because they do not want to answer that specific question or because they accidentally did not provide an answer.

Missing ImpactEd data

From our understanding, ImpactEd data can be missing for an individual for three main reasons:

1. **School not in ImpactEd:** The school is not set-up in ImpactEd, either because they did not want to sign-up or because they were not able to get set-up, for whatever reason.
2. **School does not input data:** The school is set-up, but has not inputted the data in time.
3. **Student data is missing:** The school is set-up, and has inputted data, but for whatever reason, the individual's data is missing. If we find any of this type of missing data, we will work with schools and ImpactEd to understand the reasons.

Observations with missing ImpactEd data will be dropped from the analysis and a complete case analysis run for the two ImpactEd outcomes.

Compliance

Given this is a pilot trial focused on evaluability rather than estimating impact, we do not intend to conduct any impact analysis accounting for varying levels of compliance with the intervention (e.g. IV and CACE analyses). However, we will be monitoring compliance with the intervention and reporting descriptive statistics on this (i.e. attendance rate and programme dosage/fidelity information - please see above).

Intra-cluster correlations (ICCs)

Not applicable as randomisation is conducted at the individual level.

Presentation of outcomes

The primary and secondary analyses, including the subgroup analysis, will convert effect sizes into Hedges' g (as recommended by YEF) using the following transformation:

$$Hedges'\ g\ =\ \frac{M_1 - M_2}{SD^*_{pooled}}$$

Where M_1 - M_2 is the difference in mean test scores between the treatment and control group, and SD^*_{pooled} is the pooled unconditional standard deviation.

We will report 95% confidence intervals for the effect sizes of all primary and secondary outcome analyses. The coefficients will also be reported together with one to three stars to show confidence of the estimate at different standard levels of significance (1%, 5%, 10%).