STATISTICAL ANALYSIS PLAN

Future Men's Boys Development Programme. A randomised controlled trial efficacy study

Cordis Bright

Principal investigators: Dr Stephen Boxford, Emma Andersen, Professor Darrick Joliffe, Keiran Manners, Camilla Antrobus

YOUTH ENDOWMENT FUND

Evaluation of Future Men's Boys Development Programme

Statistical analysis plan

Evaluating institution: Cordis Bright

Principal investigator(s): Dr Stephen Boxford, Emma Andersen, Professor Darrick Jolliffe, Camilla Antrobus, Keiran Manners

Project title	Future Men's Boys Development Programme. A randomised controlled trial efficacy study.
Developer (Institution)	Future Men
Evaluator (Institution)	Cordis Bright
Principal investigator(s)	Dr Stephen Boxford, Emma Andersen, Professor Darrick Joliffe, Keiran Manners, Camilla Antrobus
SAP author(s)	Emma Andersen, Professor Darrick Joliffe, Dr Stephen Boxford
Trial design	Two-armed parallel randomised controlled trial with random allocation at the young person level
Trial type	Efficacy study
Evaluation setting	Six secondary schools in South London
Target group	Boys aged 11 to 16 who are at risk of disengagement from school, exclusion or poor outcomes due to known adversity factors.
Pupil age range and Key Stage	Aged 11 to 16, Key Stages 3 and 4.
Number of participants	480 boys from six secondary schools

Primary outcome and data source	School engagement, measured by the School Connectedness Questionnaire (Marsh and Randolph, 2020).
Secondary outcome and data source	Relationships with peers, measured by the Strengths and Difficulties peer-relationship problems subscale (Goodman, 2005) Emotional symptoms, based on the Strengths and Difficulties emotional symptoms subscale (Goodman, 2005) Behavioural difficulties measured by the Strengths and Difficulties externalising behaviours score (Goodman, 2005) Relationships with teachers, measured by the School Connectedness Teacher Bonding and Attachment subscale (Marsh and Randolph, 2020). Relationship between project co-ordinator and young person, measured by the Social Support and Rejection Scale (Roffman et al., 2000).

1 SAP version history

Version	Date	Changes made and reason for revision
1.1 [*latest*]	14 March 2025	Update made to the final sample size in Section 5 following completion of baseline data collection in February 2025.
1.0 [*original*]	July 2024	

2 Table of contents

3 Introduction

This is the statistical analysis plan for an efficacy study, two-arm parallel randomised control trial (RCT) evaluation and implementation and process evaluation (IPE) of Future Men's Boys Development Programme (BDP). The BDP will be delivered across six schools in South London between September 2023 and July 2025 i.e. across two academic years. The efficacy study will run until March 2026.

The BDP is a targeted, manualised, social and emotional learning programme delivered across 12 one-to-one, 50 to 60-minute-long sessions. It aims to develop the social and emotional capacity and skills of boys in Years 7-11 who are at risk of exclusion and disengagement from schools, to improve school engagement and reduce the likelihood of exclusion.

The evaluation team are collecting a range of data throughout the trial. This includes:

- Self-reported outcomes measures, including the School Connectedness Questionnaire, the Strengths and Difficulties Questionnaire, and the Social Support and Rejection Scale.
- Demographic data provided by school partners, including age, sex, Free School Meal (FSM) eligibility, special educational needs and disabilities (SEND) data, English as an Additional Language (EAL) and care status.
- Activity and monitoring data collected by the Future Men delivery team.

This document sets out our planned analysis for each data type in more detail. The rest of this document is structured in the following way:

- **Section 4: Trial design** sets out the research questions, key outcomes and measures, and randomisation approach for the trial.
- **Section 5: Sample size calculations** presents the power calculations for the trial.
- **Section 6: Analysis** sets out the approach to analysis of primary outcomes, secondary outcomes, subgroup analysis, longitudinal analysis, further and interim analysis, missing data, compliance and outcomes presentation.

4 Trial design

4.1 Overview

The efficacy study is a two-arm, parallel randomised control trial (RCT). All young people who are referred into the project, who meet the eligibility criteria and who consent to be part of the evaluation will be allocated at random to a treatment or control group on a 1:1 basis.

Randomisation is conducted at the individual level, with stratification at the school level. Table 1 below presents an overview of the trial design.

Table 1: Trial design overview

Trial design, including number of arms		Two-arm parallel efficacy randomised controlled trial with random allocation at the young person level
Unit of randomisation		Individual young person
Stratification variables		Secondary school
Number of participants		480, i.e. 240 in the treatment group and 240 in the control group.
Primary outcome	variable	School engagement
	measure (instrument, scale, source)	School Connectedness Questionnaire (Marsh and Randolph, 2020).
Secondary outcome(s)	variable(s)	Emotional symptoms; Relationships with peers; Behaviour difficulties; Relationships with teachers; Relationship between project co-ordinator and young person.
	measure(s) (instrument, scale, source)	Relationships with peers, measured by the Strengths and Difficulties peer-relationship problems subscale (Goodman, 2005) Emotional symptoms, measured by the Strengths and Difficulties emotional symptoms subscale (Goodman, 2005) Behavioural difficulties, measured by the Strengths and Difficulties externalising behaviours score (Goodman, 2005)

		Relationships with teachers, measured by the School Connectedness Teacher Bonding and Attachment subscale (Marsh and Randolph, 2020). Relationship between project co-ordinator and young person, measured by the Social Support and Rejection Scale (Roffman et al., 2000).
Baseline for primary outcome	variable	Self-report measure of school engagement
	measure (instrument, scale, source)	School Connectedness Questionnaire (Marsh and Randolph, 2020).
Baseline for secondary outcome	variable	Relationships with peers; Emotional symptoms; Behavioural difficulties; Relationships with teachers[1].
	measure (instrument, scale, source)	Relationships with peers, measured by the Strengths and Difficulties peer-relationship problems subscale (Goodman, 2005) Emotional symptoms, based on the Strengths and Difficulties emotional symptoms subscale (Goodman, 2005) Behavioural difficulties, measured by the Strengths and Difficulties externalising behaviours score (Goodman, 2005) Relationships with teachers, measured by the School Connectedness Teacher Bonding and Attachment subscale (Marsh and Randolph, 2020).

[1] Please note at baseline, we will not be using the Social Support and Rejection Scale as the young person will not have worked with the project co-ordinator prior to the BDP.

4.2 Research questions

The key research question of the efficacy study is:

> *'Does a targeted, social-emotional learning programme for boys at risk of disengagement and exclusion improve school engagement in comparison to business as usual?'*

Additional research questions are:

1. **Delivery**: Can the BDP work under ideal circumstances?
2. **Impact**: a) What is the impact of the BDP? b) Do different sub-groups of young people have different outcomes, e.g. those from minoritized/marginalised groups?
3. **Unintended consequences**: Does the BDP have any unintentional consequences? If so, what are these? Do different groups of young people experience these differently?
4. **Iatrogenic effects**: Are there any serious negative effects attributed to the BDP on any outcomes?
5. **Mechanisms**: Which factors contribute most to the observed outcomes?

4.3 Outcomes

The primary outcome measure for the evaluation is school engagement measured by the School Connectedness Questionnaire (Marsh and Randolph, 2020). This measure was selected from the YEF outcomes framework in collaboration with YEF and Future Men colleagues.[2]

The secondary outcomes are that boys:

- Get along better with their peers, measured by the Strengths and Difficulties Questionnaire (SDQ) peer relationship problems sub-scale.
- Get along better with their teachers, measured by the SCQ teacher bonding and attachment subscale.
- Reduced behavioural difficulties, measured by the SDQ externalising behaviours score.

[2] For more information see: https://youthendowmentfund.org.uk/outcomes/.

- Reduction in emotional symptoms, measured by the SDQ emotional symptoms sub-scale.
- Have positive relationships with their project co-ordinator (control group) / significant adult (treatment group), measured by the Social Support and Rejection Scale total score.

Data for all measures is collected directly from boys by Future Men's research assistants and/or a project co-ordinator who is not delivering support using an online survey software. This is administered at baseline, 12 weeks and 24 weeks post-randomisation.

5 Sample size calculations

5.1 Overview

This section sets out:

- Power calculations for the whole cohort, which were conducted a priori as part of the initial trial design and co-design period.
- Power calculations for the subgroup of boys who are eligible for Free School Meals (FSM), which were conducted once randomisation had begun.
- An updated final sample size as of March 2025, following the completion of baseline data collection.

These calculations are summarised in Table 2 below and explained in more detail throughout this section.

Update: March 2025

Table 2 was updated in March 2025 following the completion of baseline data collection. It now sets out both the a priori power calculations completed as part of the trial design (protocol), and the final randomisation sample and MDES for the whole cohort and FSM subgroup following baseline data completion (randomisation). This shows that the trial has achieved its target baseline sample size. These calculations are explained in more detail in call out boxes in sections 5.2 and 5.3 below.

Table 2: Sample size calculations

		Protocol		Randomisation	
		Overall	FSM	Overall	FSM
Minimum Detectable Effect Size (MDES)		0.195	0.211	0.195	0.247
Pre-test/ post-test correlations	level 1 (participant)	0.7	0.7	0.7	0.7
	level 2 (cluster)	N/A	N/A	N/A	N/A
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Number of participants	Intervention	211	180	212	132
	Control	211	180	212	132
	Total	**422**	**360**	**424**	**264**

5.2 Whole cohort

We have determined the overall sample size for the trial a priori, in line with YEF guidance. These calculations suggest that a final sample size of 422 (211 per group) would ensure that the efficacy study is sufficiently powered to detect a statistically significant result if it exists (power = 0.80, two tailed, $p < 0.05$). Please note that our planned sample size for the trial is 480 boys, in line with Future Men's capacity to deliver the intervention. This provides buffer for attrition throughout the trial of 12% (58 boys). However, the final sample for the trial will not be finalised until randomisation is complete in April 2025. We will update Table 2 with

the final figures once they are known. Our approach to estimating the sample size for this efficacy study is conservative and has been influenced by the following:

- **YEF guidance.** YEF guidance suggests that efficacy study RCTs should have a Minimum Detectable Effect Size (MDES) of 0.20.
- **Available data on the School Connectedness Questionnaire.** The primary outcome measure for the study is school engagement, as measured by Marsh and Randolph's (2020) School Connectedness Questionnaire. In their validation study, Marsh and Randolph (2020) reported mean average scores, standard deviations and skew for each of the three SCQ sub-scales across two samples – one for a sample of young people in general education, and one for a sample for young people experiencing emotional and behavioural disorders. Total scores and the associated standard deviation were not reported in Marsh and Randolph's (2020) paper, and data on pre-test and post-test correlations has also not been published. As such, we have drawn from available data on similar measures (see below).
- **Available data on similar measures.** As described above, pre-test post-test correlations for the SCQ have not been published. Therefore, we draw from available data on the Student Engagement in Schools Questionnaire (Lam et al., 2014, p.38). For this scale, six month test-retest results were found to be 0.73 for the full-scale. In line with our conservative approach, we have therefore suggested a pre-test/post-test correlation of 0.7.

Table 2 at the start of this section presents power calculations which have been conducted in line with the above approach. These power calculations suggest that a sample size of 211 in each group (i.e. 422 in total) will be sufficiently powered to detect a statistically significant result if it exists (power = 0.80, two tailed, MDES = 0.195).[3]

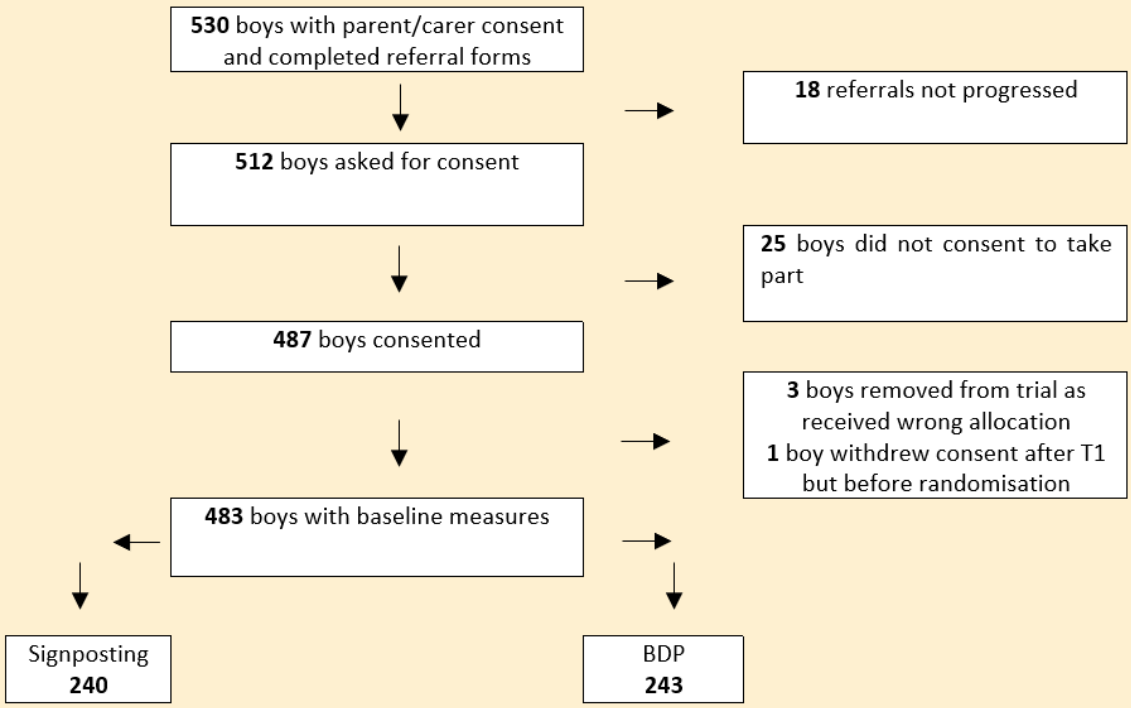
Future Men's delivery capacity enables them to deliver the BDP intervention to a maximum of 240 boys over the duration of the study. With a control group allocated as close as is feasible to a one-to-one ratio, this means a total sample size of 480 boys is achievable under existing plans. Based on the power calculation above, this means even with attrition or non-completion of questionnaires for 38 boys (12%, which is slightly above YEF's desired target of no more than 10% attrition) the sample should be able to detect statistically significant differences if they exist.

[3] These power calculations have been conducted using PowerUp. Available here: https://www.causalevaluation.org/power-analysis.html [Last accessed 17/06/2024].

Update: March 2025

Baseline data collection was completed in February 2025, with 483 baseline surveys collected (i.e, 3 more than our target of 480). Of these, 243 boys are in the BDP group, and 240 are in the signposting group, which is in line with our 1:1 allocation ratio target. The CONSORT diagram below presents this in more detail.

Figure 1: CONSORT diagram



Maintaining the above assumptions of a 12% attrition rate, this suggests that the trial will have a final sample of 425 questionnaires at T2. Rounded down to 424 (i.e. 212 in each group), power calculations suggest that this will result in a final MDES of 0.195.

5.3 Free School Meals subgroup

In line with EEF/YEF requirements, we have also conducted power calculations relating to the subgroup of boys who are eligible for Free School Meals (FSM). These calculations were conducted as part of the finalisation of this statistical analysis plan, i.e. once the trial had begun.

We have estimated that 75% of the overall cohort will be eligible for Free School Meals. This is based on the available demographic data for the boys who have been onboarded to the trial as of 4th January 2024. This shows that 75% (87 boys, n=116) are in receipt of FSM. Assuming that this trend continues, we estimate that 360 boys (i.e., 75% of the overall sample

of 480) will be eligible for FSM. This translates to 180 boys in the treatment group, and 180 boys in the control group. However, we will revisit this once baseline data collection is complete and the characteristics of the entire cohort are known.

Maintaining the assumptions outlined above (power = 0.80, two tailed, $p < 0.05$), a sample size of 360 will be powered to detect an MDES of 0.211 for this subgroup if it exists. Section 6.4 provides more information on our approach to subgroup analyses for the FSM cohort.[4]

Update: March 2025

Of the 483 boys with a baseline survey, 301 boys (62%) were reported to be eligible for free school meals by school partners. Of these, 145 boys (48%) are in the BDP group, and 156 boys (52%) are in the signposting group.

Maintaining the above assumptions of 12% attrition rate, this suggests that the trial will have a final sample of 264 questionnaires at T2 for boys who are eligible for Free School Meals. Power calculations suggest that this will result in a final MDES of 0.247 for this subgroup.

6 Analysis

6.1 Overview

This section presents a draft analytical approach for the efficacy study. This is because randomisation is still being undertaken and baseline data is not complete. As such, we currently do not know the structure of the baseline data, and this should be considered when interpreting the following analysis plan. Once all baseline data has been collected, we will revisit and finalise the analysis plan. This will be specified and documented before follow-up outcomes data has been collected, and will also include a baseline CONSORT diagram.

All analysis will be conducted on an intent-to-treat basis in line with YEF and EEF guidance. This means that participants will be analysed according to their allocation, regardless of whether they received the BDP or not. This provides the most conservative estimate of impact, as it captures the impact of offering the intervention for those who do and do not comply, and helps to fully preserve the benefits of randomisation (Torgerson and Torgerson, 2008). This approach is particularly relevant for policymakers and commissioners, i.e. those who may roll out an intervention but do not have control over take-up across the system.

[4] These power calculations have been conducted using PowerUp. Available here: https://www.causalevaluation.org/power-analysis.html [Last accessed 17/06/2024].

The analytical approach has been developed a priori and will be conducted in SPSS. The syntax for all analysis will be provided once it has been developed after all data has been collected.

6.2 Primary outcome analysis

The primary outcome is an improvement in school engagement at the individual level, as measured by the School Connectedness Questionnaire (SCQ). All young people will have completed the SCQ before randomisation, and at around 12 weeks and 24 weeks post randomisation. The primary outcome time point is 12 weeks post-randomisation.

About the School Connectedness Questionnaire (Marsh and Randolph, 2020)

The SCQ contains 10 items. Each item is marked either Not True, Somewhat True, or True, which are assigned a score of 1 to 3 respectively. It consists of three subscales:

1. Teacher bonding and attachment (3 items);
2. Peer bonding and attachment (4 items); and
3. School engagement (3 items).

The score for each subscale is the sum of its items. The primary outcome measure for this trial is the sum of the three subscale scores, and analysis of each subscale score will also be presented separately.

There is considerable debate about best practice when it comes to the analysis of data from RCTs. For example, Twisk et al. (2018) advocate for utilising longitudinal analysis of covariance or a repeated measures analysis without the treatment variable, but with the interaction between treatment and time in the model controlled for. They argue that failure to control for baseline differences in outcomes between the groups can lead to biased treatment estimates. Alternatively, others have cautioned against this approach (Sen, 2013).

Based on our understanding of the literature, our analysis of impact on the primary outcome will be conducted using a fixed effects analysis of co-variance (ANCOVA) model, which accounts for the stratification factor used in randomisation.

We will use the following model:

$$Y_i = \beta_0 + \beta_1 Baseline_i + \beta_2 Group_i + \delta School_k + \varepsilon_i \, \epsilon \, N(0, \sigma^2)$$

For $i = 1, \dots, n$ pupils per school, and $k = 1, \dots, K$ schools.

Where:

- Y_i is the school engagement subscale score as measured by the SCQ.

- $Baseline_i$ is the baseline outcome measure of the school engagement subscale for pupil i.
- $Group_i$ is a dummy variable for allocation group, i.e. 1 for the treatment group and 0 for the control group.
- β_2 is the average treatment effect, i.e. the primary parameter of interest for the trial.
- $School_k$ is a vector of $K-1$ binary school dummy variables.
- ε_i is the error/residual.
- σ^2 is the variance.

This analysis is designed to evaluate the differences in school engagement between boys in the BDP group, and boys in the business-as-usual group. Specifying the model upfront will ensure analysis avoids the "fishing problem" and the "curse of dimensionality" (Humphreys et al 2013; Hayes 2011).

We will run robustness checks to assess the underlying assumptions for ANCOVA. This will include assessing normality of the data using histograms and K-S tests. If the data does not meet these assumptions, we will run a non-parametric ANCOVA analysis. This will be determined once all data has been collected.

In line with the findings from Marsh and Randolph's (2020) original validation paper for the SCQ, it is possible that the baseline and outcome variables may be positively skewed. Skew will be assessed using the traditional criteria based on their distribution (i.e., skews of greater or equal to 1.0 or less than or equal to -1.0). Arguably, it is more desirable to use the appropriate modelling of non-normally distributed variables (e.g., Akram et al., 2023), than it is to transform the data.

6.3 Secondary outcomes analysis

Our approach to analysis of secondary outcomes will mirror the approach outlined above for primary outcome analysis.

The secondary outcomes are that boys:

- Get along better with their peers, measured by the Strengths and Difficulties Questionnaire (SDQ) peer relationship problems sub-scale.
- Get along better with their teachers, measured by the SCQ teacher bonding and attachment subscale.
- Reduced behavioural difficulties, measured by the SDQ externalising behaviours score.

- Reduction in emotional symptoms, measured by the SDQ emotional symptoms sub-scale.
- Have positive relationships with their project co-ordinator (treatment group) / significant adult (control group), measured by the Social Support and Rejection Scale.

About the Strengths and Difficulties Questionnaire (Goodman 1997).

The SDQ is a 25 item questionnaire measuring behaviours, emotions and relationships for 4 to 17 year olds. It contains 5 subscales:

1. Emotional symptoms.
2. Conduct problems.
3. Hyperactivity/inattention.
4. Peer problems.
5. Prosocial behaviour.

Each item is scored on a 3 point Likert scale from 0 to 2, such that the scores for each subscale ranges from 0 to 10. For the prosocial values subscale high scores are desirable (e.g., greater prosocial values), but for the other subscales (e.g., emotional symptoms subscale, conduct problems subscale, peer relationship subscale) high scores are not desirable (e.g., greater emotional problems, greater conduct problems, poorer peer relationships). The total difficulties score ranges from 0 to 40 and is the sum of subscales 1) to 4) above. The externalising behaviours score ranges from 0 to 20 and is the sum of the conduct and hyperactivity scores. Items from the impact supplement on overall distress and impairment generate an impact score, which ranges from 0 to 10.[5]

For each secondary outcome, we will conduct analysis of impact on the measures outlined above. This will be conducted using the same model specified for the primary outcomes measure in section 6.2 above, which will include the baseline measurement for the respective secondary outcome variable. In addition, as discussed in section 6.2 above, we will also run this analysis on the three individual subscales of the SCQ.

This analysis is designed to evaluate the differences in peer relationships, teacher relationships, emotional symptoms, behaviour, and positive relationships between the young

[5] Further information about the SDQ can be found on the YEF website here: https://youthendowmentfund.org.uk/wp-content/uploads/2022/04/18.-YEF-SDQ-guidance-April-2022.pdf

person and their project co-ordinator between those in the BDP group, and those in the business-as-usual group.

We will also run robustness checks to assess the underlying assumptions for ANCOVA for each secondary outcome measure. This will include assessing normality of the data using histograms and K-S tests. If the data does not meet these assumptions, we will run a non-parametric ANCOVA analysis. This will be determined once all data has been collected.

This analysis will be conducted in SPSS.

6.4 Subgroup analyses

The subgroup analyses we plan to undertake are likely to be exploratory in nature. Before undertaking sub-group analyses we will assess whether these would be sufficiently powered based on the data we have collected. If power calculations show that this analysis is underpowered, the analysis will be reported as exploratory, and caveated that results should be interpreted with caution.

We will assess the presence of heterogenous treatment effects in line with race equity, equality, diversity and inclusion considerations. There is limited evidence about the effectiveness of school-based interventions which aim to reduce exclusions for those from racially minoritised backgrounds. As such, we will explore whether the BDP was equally effective for those from racially minoritised backgrounds compared to those from White British backgrounds. This would likely be an underpowered analysis so caution should be applied when interpreting the results.

We will conduct this analysis by exploring the presence of interaction effects between ethnicity and treatment allocation. We will report the estimated differences across subgroups with the respective confidence intervals. This would use the following model:

$$Y_i = \beta_0 + \beta_1 Baseline_i + \beta_2 Group_i + \gamma School_k + \delta Ethnicity_{ij} + \theta Ethnicity_{ij} * Group_i + \varepsilon_i \, \epsilon \, N(0, \sigma^2)$$

For $i = 1, \dots, n$ pupils per school, with $j = 1, \dots, J$ ethnicities, at $k = 1, \dots, K$ schools

This model uses the same variables as the model set out in section 6.2. In addition:

- $Ethnicity_{ij}$ is a vector of binary dummy variables for $J - 1$ ethnicities.
- θ is a vector of parameters indicating the existence of heterogenous treatment effects by ethnicity. The total treatment effect for boys in each ethnicity grouping will be $\beta_2 + \theta$.

We will report both the point estimates and confidence intervals for θ. If θ indicates that the BDP is differentially impactful for different ethnicities, we will also consider re-running the model in section 6.2 for each sub-group separately (i.e. for boys from White, Asian, Black, Mixed and Multiple ethnic backgrounds, and Other ethnic backgrounds). If the two treatment effects are similar, this will strengthen the findings from this exploratory analysis.

We will also conduct exploratory subgroup analyses on Free School Meal (FSM) eligibility. We will include power calculations for this analysis in the final report, once the sample size for this subgroup is confirmed. However, this analysis will likely be underpowered, so results should be interpreted with caution. We will take the same approach to this analysis as outlined for ethnicity above.

In addition, based on assessment of statistical power, we may explore subgroup analyses for English as an Additional Language (EAL), and special educational needs and disability status (SEND). This analysis will be conducted on demographic data provided by school partners, and will also take the same approach as outlined for ethnicity above.

6.5 Further analyses

We will conduct the following exploratory further analysis:

- **The impact of positive relationships.** We will evaluate the extent to which positive relationships between the young person and project co-ordinator (treatment group) or significant adult (control group) influenced the primary outcome over and above the impact of the BDP through the SSRS. We are proposing conducting this analysis because the theory of change suggests that the key mechanism of change for the BDP is that it has its effect through an increase in positive relationships with a trusted adult. This will take a mediation analysis approach, i.e. we will estimate the direct and indirect effects, following the approach outlined in Gunzler et al. (2013).

About the Social Support and Rejection Scale

The SSRS has 4 dimensions: Feels valued, trust, mentoring, and negativity. Each item is scored from 1 (never) to 5 (always). Each subscale score is the average of items that make up the subscale. Higher scores on the negativity scale reflect higher levels of stress and negativity within the relationship. For the overall scoring of the scale a high score represents a positive relationship.

- **Improvements to school attendance and reductions to exclusion as measured by school administrative data.** If we are able to access school administrative data which is of sufficient high quality, consistency and comparability across schools, we will explore whether there is a difference in attendance and exclusions between those in

the BDP group and those in the business-as-usual group. This will follow the same model as set out for the primary and secondary outcomes analysis.

- **Dosage and fidelity.** Any analysis relating to dosage and fidelity will be exploratory in nature. If power calculations suggest that the analysis will be sufficiently powered, we will explore the association between the level of dosage and the impact on the SCQ. This data will be captured by monitoring data on number of sessions received collected by the Future Men delivery team. This analysis will be conducted using a general linear model assuming normality, or a generalized linear model. It will answer questions such as: does attending 8 or more BDP sessions result in a similar impact as attending all 12 sessions?

6.6 Interim analyses and stopping rules

After the first cohort of baseline data collection we will analyse the completeness, reliability and validity of outcomes questionnaires (including the outcomes measures described above). We will do this by exploring:

- Percentages of scale item completeness.
- Outcome measure means, standard deviations and skew.
- Cronbach Alpha testing for scale reliability.

This analysis will not include a comparison between control group and treatment group data nor analysis of impact. We will review and discuss these findings with Future Men and YEF to provide reassurance that data collection is proceeding well. If there are concerns, we will suggest and discuss solutions with YEF and Future Men.

We will continue to monitor data quality looking at scale completeness, means, standard deviations and skew throughout the trial for internal purposes to ensure that data collection proceeds smoothly. If on review we have concerns, we will raise this with YEF and Future Men as appropriate.

The trial will stop if the BDP is unable to recruit a sufficient number of participants. Recruitment rates will be regularly monitored against modelled target rates, and reviewed as part of project group meetings. Any decisions about stopping will be made in discussion with YEF and Future Men colleagues.

The Future Men project team will also be responsible for safeguarding of participants. They will report any serious adverse events overall and by trial arm. The trial will stop if Future Men, YEF and Cordis Bright decide that the BDP is unsafe for participants.

6.7 Longitudinal follow-up analyses

As discussed in section 6.2, boys in both groups will complete outcomes measures at both 12 and 24 weeks post-randomisation. We will conduct longitudinal analysis of covariance to assess the extent to which the impact of the BDP on school engagement has been sustained. This will be conducted in line with the longitudinal ANCOVA approach recommended in Twisk et al. (2018), and will take the following form:

$$Y_i = \beta_0 + \beta_1 Baseline_i + \beta_2 Group_i + \delta School_k + \beta_4 Time_{it} + \beta_5 Group_i * Time_{it} + \varepsilon_i \in N(0, \sigma^2)$$

Where:

- β_2 reflects the treatment effect at 12 weeks post randomisation.
- $Time_{it}$ is a dummy variable for 24 weeks post randomisation.
- $\beta_2 + \beta_5$ is the treatment effect at 24 weeks post randomisation.

We will report both point estimates and confidence intervals for impact estimates at both 12 and 24 weeks post randomisation. Differences between the two and their implications will be discussed in the final report.

6.8 Imbalance at baseline

If randomisation has been successful, both treatment and business-as-usual groups should be equivalent at baseline. As such, any imbalance will have occurred by chance. To check for and monitor imbalance at baseline, we will produce a table of descriptive characteristics for all young people at baseline. We will also produce an equivalent table for those included in the final analysis sample, to check whether any attrition experienced throughout the trial may have introduced an imbalance. These descriptive characteristics will include:

- Age.
- Sex.
- Ethnicity.
- Free school meal eligibility.
- English as an additional language.
- SEND status.
- Baseline SCQ scores.

- Baseline SDQ scores.

Data on pupil characteristics is provided by school partners through the initial referral form. We will present a cross-tabulation of counts and percentages for each category above against allocation group. For continuous variables, we will present the means and standard deviation. This analysis will be used to inform our understanding of the extent to which our initial sample was balanced across the two groups, and whether any attrition experienced throughout the trial has introduced an imbalance. We will discuss any differences and their implications in the final report.

6.9 Missing data

Throughout the trial, the evaluation team will work closely with the Future Men team to support the collection of high quality and complete data for all boys. However, missing data may occur due to either item non-completion or sample attrition (i.e. boys who do not complete either T2 or T3 questionnaires). We will assess both the extent of missingness and patterns of missingness in the data. In line with YEF guidance we will report on both: (1) the proportion of missing data in the trial, and (2) the extent and pattern of missingness in the data. This will involve analysis of whether data is missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR and MAR mean that complete cases are unlikely to be biased subsequent to adjustment but may be underpowered, while MNAR suggests that structural bias has been introduced to the sample.

We will attempt to establish the missing mechanism (i.e. which variables in the data are predictive of non-response) through logistic regression models. This will model the presence of missing outcomes data with additional information that may be predictive of missingness. We will conduct this analysis in line with the flow chart in Figure 1 in the YEF analysis guidance.[6] This outlines the following approach:

- If the prevalence of missing data is less than 5%, no further action is required as complete case analysis is unlikely to be biased.
- If outcomes data is MAR conditional on co-variates, we will include these co-variates in our primary analysis model and discuss the implications in full.
- If a covariate is MAR conditional on other covariates, we will conduct multiple imputation (MI). Treatment effects from the MI analysis will be reported in addition to

[6] Available here, page 15: https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf [Last accessed 12/12/2023]

estimates from the model outlined in section 6.2. Any differences between the two and their implications will be discussed in full.

- If missing data cannot be fully explained by the other variables in the dataset, data is likely to be MNAR. In this scenario we will conduct a sensitivity analysis alongside the primary impact analyses.

We will only conduct the above analyses for the primary outcome analysis, as all secondary outcomes analysis, subgroup analysis and further analysis is tentative and exploratory in nature.

Unfortunately, there is no universally agreed approach to analysis in the event of item non-completion. In the event that a high proportion of cases would be excluded due to low rates of item non-completion (for example, if most boys miss a small number of items), our approach to missing data will balance considerations around data integrity with maximising statistical power. In this scenario, we would consider using statistical techniques to impute missing items. We will finalise and agree our approach to this for the final draft of the Statistical Analysis Plan in line with YEF guidance, i.e. once baseline data collection is complete and we have a greater understanding of the structure of the data.

6.10 Compliance

As outlined in section 6.1, all analysis will be conducted on an intent-to-treat basis. This means that overall compliance for the purposes of the efficacy study will be met when young people have been randomised and allocated into the treatment or control group.

However, we acknowledge that intent-to-treat analysis may underestimate the efficacy of the intervention if some boys in either trial arm do not adhere to their assigned treatment. To examine this, we will conduct Complier Average Causal Effect (CACE) analysis, which will indicate treatment effects amongst those who comply with the intervention. However, any analysis of treatment effects in the presence of non-compliance will also be exploratory. This will be estimated using two stage least squares (2SLS) regression (Gerber and Green, 2012), which uses the following two stages:

1. The first stage will model the compliance variable (i.e. number of sessions) using the same explanatory variables used for the primary analysis. This will be a logistic regression model used to generate predicted compliance.

2. The second stage models will use predicted compliance in place of the allocation group variable in the ITT primary analysis specified in section 6.2 to generate the CACE estimates.

We will report the results from the first stage of the 2SLS, along with the correlation between the instrument and endogenous variable and the associated F-test. Interpretations of the CACE estimates will be provided in the final report.

6.11 Intra-cluster correlations (ICCs)

This is not a clustered randomised controlled trial. As such, ICCs will not be calculated.

6.12 Presentation of outcomes

Effect sizes will be calculated using Hedges' g, as specified in the following equation:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{sd_{pooled}}$$

Where:

- $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the ANCOVA difference in means between the treatment and control groups adjusted for baseline outcomes measures and school, as specified in the primary outcomes model.
- sd_{pooled} is the unconditional pooled standard deviation of the two groups.[7]

With a sample of greater than 20 there is limited difference with Cohen's d. However, if the standard deviations between the treatment and comparison group are different, we would propose to use Glass' delta, which only uses the control group's standard deviation (Lipsey & Wilson, 2001).

We will report the statistical uncertainty associated with the above effect sizes through both the confidence intervals and the p value. Confidence intervals will be calculated using the following formula:

$$g \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) g_{se}$$

Where:

- Φ^{-1} is the percent point function of the normal distribution.

[7] $sd_{pooled} = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1+n_2-2}}$, where n_1 and n_2 are the sample size for groups 1 and 2 respectively, and sd_1 and sd_2 are the standard deviations of group 1 and group 2 respectively.

- g_{se} is the standard error of the g statistic (noted as ES above). [8]

All estimations and their statistical uncertainty will be reported, and the implications of both the point estimates and confidence intervals will be set out. In addition, all reporting will consider findings in light of the existing evidence base. This will be triangulated with the evidence collected from the implementation and process evaluation on the quality and context of delivery, the existence of theoretical causal mechanisms, and the experiences and perspectives of boys, practitioners and wider school stakeholders who participate in semi-structured interviews.

7 References

Akram, M., Cerin, E., Lamb, K. E., & White, S. R., 2023. Modelling count, bounded and skewed continuous outcomes in physical activity research: beyond linear regression models. *International Journal of Behavioural Nutrition and Physical Activity*, *20*(1), 1-11.

Chow S-C, Shao J, Wang H. Sample size calculations in clinical research. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2008. Section 3.2.1, page 58.

Education Endowment Foundation, 2022. *Statistical analysis guidance for EEF evaluations.* Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1702283680 [Last accessed 11/12/23].

Gerber and Green, 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Goodman, R., 2005. *The Strengths and Difficulties Questionnaire.* Available at: https://www.sdqinfo.org/a0.html [Last accessed 11/12/23].

Gunzler D, Chen T, Wu P, Zhang H. Introduction to mediation analysis with structural equation modeling. Shanghai Arch Psychiatry. 2013 Dec;25(6):390-4. doi: 10.3969/j.issn.1002-0829.2013.06.009. PMID: 24991183; PMCID: PMC4054581.

Humphreys, M., De la Sierra, R.S. and Van der Windt, P., 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. Political Analysis, 21(1), pp.1-20.

[8] $g_{se} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{g^2}{2(n_1+n_2)}}$

Hayes, B., 2011. An adventure in the Nth dimension. American Scientist, 99(6), pp.442-446.

Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing clinical research : an epidemiologic approach. 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2013. Appendix 6A, page 73.

Lam, S. F., Jimerson, S., Wong, B. P. H., Kikas, E., Shin, H., Veiga, F. H., Hatzichristou, C., Polychroni, F., Cefai, C., Negovan, V., Stanculescu, E., Yang, H., Liu, Y., Basnett, J., Duck, R., Farrell, P., Nelson, B., & Zollneritsch, J., 2014. Understanding and measuring student engagement in school: the results of an international study from 12 countries. School psychology quarterly : the official journal of the Division of School Psychology, American Psychological Association, 29(2), 213–232.

Lipsey, M. and Wilson, D., 2021. *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Marsh, R.J. and Randolph, J.J., 2020. Preliminary development and validation of the School Connectedness Questionnaire. Child Indicators Research, 13(5), pp.1637-1654.

Roffman, J. G., Pagano, M. E., and Hirsch, B. J., 2000. *Social support and rejection scale*. Evanston, IL: Human Development and Social Policy, Northwestern University.

Senn, S., 2013. Seven myths of randomisation in clinical trials. *Statistics in medicine*, *32*(9), 1439-1450.

Torgerson, D. J., & Torgerson, C. J. (2008). Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction. London: Palgrave Macmillan.

Twisk J., Bosman, L., Hoekstra T., Rijnhart, J., Welten, M. & Heymans M., 2018. Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications*, 10, 80-85.

Youth Endowment Fund, 2021. *Analysis Guidance.* Available at: https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf [Last accessed 11/12/23].

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413