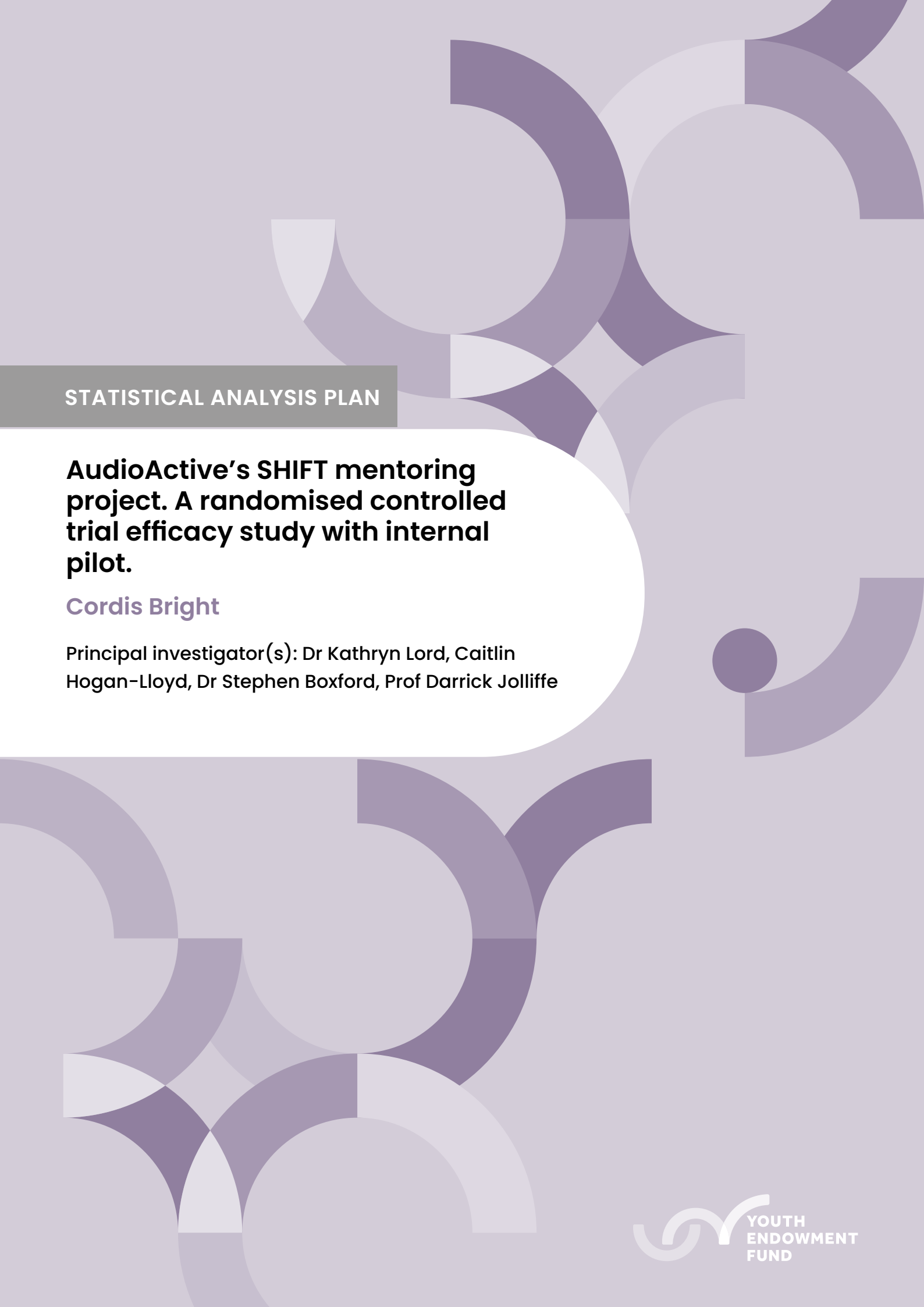




STATISTICAL ANALYSIS PLAN

AudioActive's SHIFT mentoring project. A randomised controlled trial efficacy study with internal pilot.

Cordis Bright

Principal investigator(s): Dr Kathryn Lord, Caitlin Hogan-Lloyd, Dr Stephen Boxford, Prof Darrick Jolliffe

Evaluation of the SHIFT programme

Statistical analysis plan

Evaluating institution: Cordis Bright

Principal investigator(s): Dr Kathryn Lord (PI). Co-PIs: Caitlin Hogan-Lloyd, Dr Stephen Boxford, Professor Darrick Jolliffe



YEF statistical analysis plan

Project title¹	AudioActive's SHIFT mentoring project. A randomised controlled trial efficacy study with internal pilot.
Developer (Institution)	AudioActive
Evaluator (Institution)	Cordis Bright
Principal investigator(s)	Dr Kathryn Lord, Caitlin Hogan-Lloyd, Dr Stephen Boxford, Prof Darrick Jolliffe
SAP author(s)	Dr Kathryn Lord, Caitlin Hogan-Lloyd, Dr Stephen Boxford, Prof Darrick Jolliffe, Karim Bukleb
Trial design	Two-armed parallel randomised controlled trial with random allocation at the young person level
Trial type	Efficacy study with internal pilot.
Evaluation setting	The intervention is delivered at a variety of safe spaces in the young people's community, including schools.

¹ Please make sure the title matches that in the header and that it is identified as a randomised trial as per the CONSORT requirements (CONSORT 1a).

Target group	Young people aged 11-17 at risk of, or already involved with, violence and crime across East Sussex, West Sussex, and Brighton and Hove.
Intended number of participants	586 (293 in treatment group, 293 in control group)
Primary outcome and data source	Delinquent behaviour. Measured by Self-Reported Delinquency Scale (SRDS) volume score at baseline before randomisation (T1) and at 26 weeks (+/-4 weeks) post-randomisation (T2).(see, Smith & McVie, 2003)
Secondary outcome and data source	<p>All measured at baseline before randomisation (T1) and at 26 weeks (+/-4 weeks) post-randomisation (T2).</p> <ol style="list-style-type: none"> 1. Delinquent behaviour, measured by the SRDS variety score. 2. Wellbeing, measured by the Shortened Warwick Edinburgh Mental Wellbeing Scale (S-WEMWBS) (Stewart-Brown et al., 2009). 3. Behavioural difficulties, measured by the Strength and Difficulties Questionnaire externalising score (SDQ) (Goodman, 2005). 4. Pro-social values and behaviours, measured by the SDQ pro-social sub-scale (Goodman, 2005). 5. Conduct problems, measured by the SDQ conduct problems sub-scale (Goodman, 2005). 6. Emotional problems, measured by the SDQ emotional problems sub-scale (Goodman, 2005). 7. Peer problems, measured by the SDQ peer problems sub-scale (Goodman, 2005). 8. Hyperactivity/inattention, measured by the SDQ hyperactivity/inattention sub-scale (Goodman, 2005).

1 SAP version history

Version	Date	Changes made and reason for revision
1.0	13 th October	

2 Table of contents

1	SAP version history	3
2	Table of contents	3
3	Introduction	4
4	Trial design.....	4
4.1	Trial overview	4
4.2	Research questions	6
4.3	Outcome measures	7
5	Sample size calculations	10
5.1	Protocol stage power calculations	10
5.2	Updated power calculations	12
6	Analysis.....	13
6.1	Overview.....	13
6.2	Primary outcome analysis	14
6.3	Secondary outcomes analysis	16
6.4	Subgroup analyses.....	16
6.5	Further analyses	17
6.6	Imbalance at baseline	19
6.7	Missing data	20
6.8	Compliance.....	21
6.9	Intra-cluster correlation	22
6.10	Presentation of outcomes.....	22
7	References	24

3 Introduction

This is an efficacy study statistical analysis plan (SAP) for a two-armed parallel randomised controlled trial (RCT) evaluation of AudioActive’s SHIFT programme. The efficacy study included an internal pilot trial for which data collection took place between 1st October 2024 and 31st March 2025. The trial moved to a full efficacy study in May 2025, and data collection is due to complete by 30th November 2026. Reporting will be finalised by May 2027.

The SHIFT programme is a voluntary 18 session music mentoring programme, delivered within a 26-week window. It pairs young people who are already involved in or at risk of offending with a mentor. The mentor aims to deliver weekly, 90-minute face-to-face sessions in schools, or other community settings as required.

The evaluation team, with the support of AudioActive, is collecting a range of data throughout the trial, including self-reported outcomes, demographic data, and activity and monitoring data. This document sets out our planned analysis for each type of data collected.

This plan should be read in conjunction with the study trial protocol, which is available at: <https://youthendowmentfund.org.uk/wp-content/uploads/2024/08/Shift-Evaluation-Protocol-October-2024-1.pdf>

4 Trial design

4.1 Trial overview

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

The table below presents an overview of the efficacy study trial design.

Figure 1: Trial design overview

Trial design, including number of arms	Two-armed parallel RCT with random allocation at the young person level
Unit of randomisation	Individual young person
Stratification variables (if applicable)	None

Primary outcome	Variable	Self-reported delinquent behaviour (violent and non-violent)
	measure (instrument, scale, source)	Self-Reported Delinquency Scale (SRDS) volume score at baseline before randomisation (T1) and at 26 weeks (+/-4 weeks) post-randomisation (T2). (see, Smith & McVie, 2003)
Secondary outcome(s)	variable(s)	<ol style="list-style-type: none"> 1. Delinquent behaviour. 2. Wellbeing. 3. Behavioural difficulties. 4. Positive and pro-social identity. 5. Conduct problems. 6. Emotional problems. 7. Peer problems. 8. Hyperactivity/inattention.
	measure(s) (instrument, scale, source)	<p>All measured at baseline before randomisation (T1) and at 26 weeks (+/-4 weeks) post-randomisation (T2).</p> <ol style="list-style-type: none"> 1. Delinquent behaviour measured by the SRDS variety score. 2. Wellbeing measured by the Shortened Warwick Edinburgh Mental Wellbeing Scale (S-WEMWBS) (Stewart-Brown et al., 2009). 3. Behavioural difficulties, measured by the Strengths and Difficulties Questionnaire (SDQ) externalising score (Goodman, 2005). 4. Pro-social values and behaviours, measured by the SDQ pro-social sub-scale (Goodman, 2005). 5. Conduct problems, measured by the SDQ conduct problems sub-scale (Goodman, 2005). 6. Emotional problems, measured by the SDQ emotional problems sub-scale (Goodman, 2005). 7. Peer problems, measured by the SDQ peer problems sub-scale (Goodman, 2005). 8. Hyperactivity/inattention, measured by the SDQ hyperactivity/inattention sub-scale (Goodman, 2005).
	Variable	Self-reported delinquent behaviour

Baseline for primary outcome	measure (instrument, scale, source)	SRDS volume score.
Baseline for secondary outcome	Variable	<ol style="list-style-type: none"> 1. Delinquent behaviour. 2. Wellbeing. 3. Behavioural difficulties. 4. Positive and pro-social identity. 5. Conduct problems. 6. Emotional problems. 7. Peer problems. 8. Hyperactivity/inattention.
	measure (instrument, scale, source)	<ol style="list-style-type: none"> 1. Delinquent behaviour, measured by the SRDS variety score. 2. Wellbeing, measured by the S-WEMWBS. 3. Behavioural difficulties, measured by the SDQ externalising score. 4. Pro-social values and behaviours, measured by the SDQ pro-social sub-scale. 5. Conduct problems, measured by the SDQ conduct problems sub-scale (Goodman, 2005). 6. Emotional problems, measured by the SDQ emotional problems sub-scale (Goodman, 2005). 7. Peer problems, measured by the SDQ peer problems sub-scale (Goodman, 2005). 8. Hyperactivity/inattention, measured by the SDQ hyperactivity/inattention sub-scale (Goodman, 2005).

4.2 Research questions

The primary research question of the efficacy study is:

Is a music-based mentoring programme delivered with young people aged 11-17 involved in or deemed to be at risk of involvement in violent or non-violent offending, focused on managing behaviour and developing positive relationships and a pro-social identity, an effective approach to reducing

young people's future engagement in delinquent behaviour compared to business-as-usual services?

The secondary research questions are:

1. **Delivery:** Can the SHIFT programme work under ideal circumstances?
2. **Impact:**
 - a. What is the impact of the SHIFT programme on the **secondary outcome areas** (see outcomes listed in Section 4.3 below).
3. Do different **sub-groups of young people**, including those from minority ethnic groups, those with special educational needs and disabilities (SEND), those who are children looked after (CLA) and those who are eligible for free school meals (FSM), have different outcomes?
4. **Unintended consequences:** a) Does the SHIFT programme have any unintentional consequences? If so, what are these? b) Do different groups of young people experience these differently?
5. **Iatrogenic effects:** Are there any serious negative effects that can be attributed to the SHIFT project on any outcomes?
6. **Mechanisms:** a) How does the SHIFT programme work to reduce young people's future engagement in delinquent behaviour? b) Which factors contribute most to the observed outcomes?

4.3 Outcome measures

All measures were selected in agreement with colleagues from AudioActive and YEF. Data for all outcome measures is collected directly from participants using either online survey software (SmartSurvey – see: <https://www.smartsurvey.co.uk/>) or a paper questionnaire administered at baseline and 22-30 weeks (26 weeks +/-4) post-randomisation.

Primary outcome measure

The primary outcome measure for the study is self-reported delinquent behaviour², as measured by the volume score of the SRDS at baseline/time 1 (T1) and follow up/time 2 (T2), i.e. 26 weeks (22-30 weeks i.e. 26 weeks +/-4) post-randomisation.

About the SRDS Volume Score

The SRDS is a 19-item self-reported delinquency scale developed as part of the Edinburgh Study of Youth Transitions and Crime (Smith & McVie, 2003). It covers a range of antisocial and offending behaviours, has been validated for use with young people in the UK, and has been used with those aged between 10 and 17. The volume score is calculated by converting the reported frequency for each item into a numerical score and summing these scores across items.

The SRDS has been shown to have good psychometric properties; reported internal consistency is between .87-.92 with an inter-item correlation of .19 (Fonagy et al., 2018; Humayun et al., 2017) and the measure correlates with official police arrests (89.5% - 95.2%; McAra & McVie, 2005).

More information on the subscales, psychometric properties and validity of the SRDS is available in the YEF outcomes measures database (Youth Endowment Fund, 2022b) and in the YEF core measurement guidance (Youth Endowment Fund, 2021a and 2022a).

Secondary outcome measures

The secondary outcomes, measured at baseline before randomisation (T1) and at 26 weeks (+/-4 weeks) post-randomisation (T2) that we are investigating are:

- **Variety of delinquent behaviour:** measured by the SRDS variety score.
- **Wellbeing:** measured by the S-WEMWBS
- **Behavioural difficulties:** measured by the SDQ externalising score
- **Pro-social values and behaviours:** measured by the SDQ pro-social sub-scale
- **Conduct problems:** measured by the SDQ conduct problems sub-scale

² For this study delinquent behaviour is defined as actions or conduct that violate legal, social and moral norms and are deemed harmful or disruptive to individuals, groups or society as a whole. Such behaviour may be categorised into various types based on the nature and severity of the violation, including: (1) criminal offences: Acts that breach criminal laws, such as theft, assault, fraud, and other activities punishable by law; (2) anti-social behaviour: Actions that disrupt societal norms and community standards, like vandalism, public disorder, and harassment; (3) ethical violations: Conduct that breaches moral or ethical standards, such as dishonesty, betrayal, and exploitation, often consider unacceptable.

- **Emotional problems:** measured by the SDQ emotional problems sub-scale
- **Peer problems:** measured by the SDQ peer problems sub-scale
- **Hyperactivity/inattention:** measured by the SDQ hyperactivity/inattention sub-scale

About the scales used to measure secondary outcomes

The **SRDS variety of delinquency score** uses the same 19 SRDS items as the primary outcome measure, but captures the breadth of behaviours reported, rather than frequency. It is calculated as the sum of the number of different behaviours the respondent has answered 'yes' to engaging in.

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

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

Each item is scored on a three-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 (i.e., greater prosocial values), but for all four difficulty subscales high scores are not desirable (i.e., greater conduct problems). The SDQ has been shown to have good internal consistency (Cronbach's $\alpha = .73$), cross-informant correlation (mean=0.34), and retest stability after four to six months (mean: 0.62) (Goodman, 2001).

The **S-WEMWBS** uses seven items to measure change in wellbeing. Each item is scored from 1 (none of the time) to 5 (all of the time). The S-WEMWBS is a shortened version of the full 14 item scale, and scores of the 7-item version are transformed to generate a comparable total score using a conversion table. Scores range from 7 to 35 and higher scores indicate higher positive mental wellbeing.

The SHIFT trial protocol identified quality of relationship with a mentor/case worker, measured using the Social Support and Rejection Scale (SSRS), as a secondary outcome. However, we will not analyse the SSRS as a secondary outcome in the efficacy analysis. This reflects the way in which the measure is administered and its intended role within the programme theory of change.

The SSRS is collected only at follow-up and therefore does not have an equivalent baseline measure that can be included in the ANCOVA approach used for the other secondary outcomes (see Section 6.3). In addition, the measure is not administered to equivalent samples across trial arms: all young people in the SHIFT group complete the SSRS in relation to their SHIFT mentor, whereas young people in the control group complete it only where

they identify a significant non-parental adult with whom they have a regular relationship. The resulting control-group sample is therefore selective and does not provide a directly comparable outcome measure across the full randomised sample.

For these reasons, the SSRS will instead be treated as an exploratory measure of a proposed mechanism of change. This is consistent with the more detailed description in the trial protocol, which identifies the quality of the mentoring relationship as a hypothesised mechanism through which SHIFT may affect other outcomes. Analysis of the SSRS and its relationship with the primary outcome is therefore described separately in Section 6.5.

5 Sample size calculations

5.1 Protocol stage power calculations

When drafting the original trial protocol, we estimated the sample size for the efficacy study a priori, conducting power calculations in line with YEF guidance. Figure 2 below presents the results of our power analysis at the protocol stage.

Our approach was conservative and influenced by the following:

- **YEF guidance.** YEF guidance suggests that efficacy study RCTs should have a Minimum Detectable Effect Size (MDES) of 0.20. At the protocol stage, we used a simplified prevalence-based approximation to contextualise an effect of this magnitude because we did not yet have robust information on the likely distribution of SRDS volume scores or the pre-test/post-test correlation in a comparable population. This was an approximate benchmark for the expected magnitude of effect rather than a change to the primary outcome, which remained the SRDS volume score. Following learning during the trial and the availability of more relevant information, we have refined the power calculation using an approach more closely aligned with the planned continuous-outcome analysis, as described in Section 5.2.
- **The evidence base.** The YEF mentoring toolkit (Gaffney, Jolliffe and White, 2022) suggests that similar mentoring/case worker programmes can lead to, on average, a 21% reduction on violence, a 14% reduction in all offending, and a 19% reduction in reoffending. In addition, in a meta-analysis using a random effects model ($d=.21$, 95% confidence interval, .07 to .34) of 18 studies, Jolliffe and Farrington (2008) showed that mentoring programmes make a 10-11% difference in relation to offending.³

³ Please note that this rapid evidence assessment found that mentoring was more effective in reducing reoffending when contact between mentor and mentee was greater, in smaller scale studies, and when mentoring was combined with other services and interventions.

- **Estimated SHIFT programme recruitment rates.** We also considered the monthly programme recruitment rates estimated by AudioActive based on their experience of delivering SHIFT in Brighton and Hove, Crawley and Hastings, which assumed an average monthly recruitment rate of 44 young people.
- **Pre-test/post-test correlation:** At the protocol stage, we assumed a pre-test/post-test correlation of zero. This was because we did not yet have robust evidence about the likely correlation between baseline and follow-up SRDS scores in a population similar to SHIFT participants. Although we expected that including the pre-test measure as a covariate in the impact analysis would improve precision, the likely size of this improvement was uncertain. In general, a stronger pre-test/post-test correlation reduces the MDES, because baseline scores explain some of the residual variance in post-test outcomes and therefore reduce the standard error of the impact estimate. However, in the absence of suitable evidence at the protocol stage, we did not assume any such gain in precision.

Based on these assumptions, the original protocol-stage calculation estimated that a total Time 2 sample of **586 young people** (293 in each group) would be required to detect an effect of approximately 0.20 with 80% power using a two-tailed test at the 5% significance level. At this stage, the expected effect was contextualised using an assumed reduction in offending prevalence from 30% in the control group to 20% in the intervention group, drawing on Jolliffe and Farrington's (2008) meta-analysis of mentoring programmes. This prevalence-based assumption was used as an approximate benchmark for the expected magnitude of effect in the absence of more suitable information on the distribution and longitudinal properties of the SRDS volume score.

To achieve this minimum T2 sample of 586, the original trial protocol set a recruitment target of 799 young people. Allowing for anticipated attrition of 20% following randomisation, this was expected to result in a T2 sample of approximately 639, providing a buffer above the minimum sample required. In total, 682 young people were randomised into the trial and completed baseline data during the recruitment period. The final achieved T2 sample will be confirmed following completion of data collection and the detectable effect size calculated.

Following completion of baseline data collection, we were able to refine the power calculation using assumptions more closely aligned with the planned primary analysis. The updated calculation therefore treats the SRDS volume score as a continuous outcome, incorporates an evidence-based estimate of the pre-test/post-test correlation and uses the achieved T1 sample rather than the anticipated T2 sample in line with revised YEF guidance, as described in Section 5.2.

5.2 Updated power calculations

Following completion of randomisation and baseline data collection, we recalculated the MDES using the achieved baseline sample and updated design parameters. This analysis is based on the numbers randomised and included in the sample at T1. This calculation is also presented in Figure 2. The cumulative T1 sample at the end of the recruitment period was **682**. 342 young people were allocated to the intervention group, and 340 to the control group.

We have included a **pre- post-test correlation of 0.5** within this revised power calculation. This is based on values obtained from unpublished data from an RCT using the same outcome measure and in a similar population of adolescents (Humayun et al., 2019) which we discovered after our original power analysis for the protocol. It is also based on treating the SRDS volume score as a continuous variable in the analysis. At the protocol stage, we did not have sufficiently robust evidence on the likely distribution of SRDS volume scores or the correlation between baseline and follow-up SRDS scores in a comparable population. The original power calculation therefore used a more conservative, prevalence-based approximation drawn from the mentoring literature. Once baseline recruitment was complete and more relevant evidence became available, we refined the calculation to reflect the planned primary analysis more directly by treating the SRDS volume score as continuous and incorporating an evidence-based pre-test/post-test correlation.

Our power analysis shows an MDES of 0.186, calculated based on the final randomised sample of 682 people in line with updated YEF guidance.

SPSS 25⁴ was used for calculations for the original protocol. PowerUp! (Dong, N. and Maynard, R. A., 2013) was used for calculations once all randomisation and baseline data was collected.

The final achieved T2 analysis sample will be reported in the full impact report for the efficacy study, and the achieved MDES will be recalculated based on the number of participants with valid primary outcome data at T2.

Figure 2: Sample size calculations

	Protocol	Randomisation
Minimum Detectable Effect Size (MDES)	0.20	0.186

⁴ IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

		Protocol	Randomisation
Pre-test/ post-test correlations	Level 1 (participant)	0.0	0.5
	Level 2 (cluster)	N/A	N/A
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided
Number of participants	Intervention	293	342
	Control	293	340
	Total	586	682

6 Analysis

6.1 Overview

We intend to estimate the difference between the SHIFT group and the control group for each of the primary and secondary outcomes. The analytic approach has been developed a priori and will be conducted with SPSS 25 and R.

All analysis will be conducted once delivery comes to an end in November 2026. An intent-to-treat approach to analysis will be used, in line with YEF guidance. This means that all participants will be analysed according to their allocation, regardless of whether they received SHIFT or not.

An intention-to-treat approach provides the most conservative estimate of impact, as it evaluates the impact of offering the intervention to those who do and do not comply and helps to fully capture the ‘true’ benefit (if any) of the intervention (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 of the intervention across the system.

The syntax for all analysis will be provided after reporting is complete.

6.2 Primary outcome analysis

The primary outcome is the **volume of delinquent behaviour** at the individual level as measured by the SRDS, completed before randomisation and at around 26 weeks post randomisation (+/-4 weeks). Young people are asked to report behaviours they have been involved in over the previous six months.

Questionnaires completed within 22–30 weeks post-randomisation will constitute the intended T2 assessment window. Where questionnaires are completed outside this window, they will not automatically be excluded from the primary intention-to-treat analysis, provided the outcome data are otherwise valid.

We will report the number and proportion of questionnaires completed within and outside the intended window, together with the distribution of time from randomisation to follow-up. We will also compare the baseline characteristics of participants who completed T2 within the intended window with those who completed outside it, including key demographic characteristics and baseline SRDS scores, to assess whether late completion appears systematically associated with participant characteristics. We will prioritise assessing any difference by trial group first, and will consider sensitivity analysis off the back of that analysis.

If a meaningful number or proportion of questionnaires are completed outside the specified window, particularly where there are identifiable differences between in-window and out-of-window completers, we will conduct a sensitivity analysis restricting the primary analysis to participants whose T2 questionnaire was completed within 22–30 weeks post-randomisation and compare the resulting treatment estimate with the main analysis. Where only a very small number of questionnaires fall outside the window, such that exclusion would be unlikely to materially affect the estimate, this sensitivity analysis may not be undertaken; the timing of these questionnaires will instead be reported descriptively.

There is debate about best practice in 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. They argue that adjustment for baseline outcome measures can improve statistical precision and account for chance baseline imbalances. However, others have cautioned against this approach (e.g., Senn, 2013).

Based on our understanding of such literature, our analysis of the impact of the SHIFT programme on the SRDS volume score will be conducted using a covariance (ANCOVA) model, controlling for baseline SRDS score. The ANCOVA analysis will estimate the adjusted mean difference in T2 SRDS volume scores between young people allocated to SHIFT (the treatment group) and those allocated to the control group, adjusting for baseline SRDS volume score.

The outputs from this analysis will be used to calculate the Hedges' G effect estimate (described in section 6.10).

The primary analysis will be based on observed data, including participants with valid baseline and follow-up SRDS volume scores. Multiple imputation will be considered as a sensitivity analysis, as described in Section 6.7.

We do not currently propose routinely adjusting the primary treatment-effect estimate for additional demographic characteristics, as participants were individually randomised and no stratification variables were used. Baseline characteristics, including age, ethnicity, SEND status, care status and eligibility for free school meals, will nevertheless be examined descriptively for balance between trial arms. These demographic data are collected by AudioActive through referral and monitoring data.

We will use the following model:

$$Y_i = \beta_0 + \beta_1 \text{Baseline}_i + \beta_2 \text{Group}_i + \varepsilon_i \in N(0, \sigma^2)$$

For $i = 1, \dots, n$ young people

Where:

- Y_i is the T2 (c26 weeks post-randomisation) volume score for self-reported delinquency as measured by the SRDS.
- Baseline_i is the baseline outcome measure of the SRDS volume score for young person 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.
- ε_i is the error/residual.
- σ^2 is the variance.

Specifying this model upfront will help to ensure that the analyses avoid the “fishing problem” and the “curse of dimensionality” (Humphreys et al 2013; Hayes 2011).

We will assess the assumptions relevant to ANCOVA, including linearity, homogeneity of regression slopes and the distribution and variance of residuals. We will use graphical diagnostics alongside formal checks, recognising that ANCOVA is robust to moderate departures from normality. Because SRDS volume scores are non-negative counts and may include many zeros and right-skew, if the linear model is clearly inappropriate we will use a

pre-specified generalised model suited to the observed distribution (for example, Poisson or negative binomial for count data, with a zero-inflated/hurdle specification considered if there is substantial excess zero inflation). The alternative model and rationale will be reported transparently.

Variable transformation

We will assess skew in baseline and outcome variables 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 a generalised linear model for 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 the analysis of secondary outcomes will mirror the approach outlined above for that of the primary outcome.

For each secondary outcome, we will estimate the difference between the treatment and control groups using an ANCOVA model, adjusting for the baseline value of the relevant outcome where available. The appropriateness of the ANCOVA model for each secondary outcome will be assessed using the approach described in Section 6.2.

These analyses will be conducted in SPSS 25 and R.

6.4 Subgroup analyses

All subgroup analyses will be exploratory. Before undertaking subgroup analyses, we will assess whether the available sample sizes are sufficient to support meaningful analysis.

Where individual categories contain insufficient numbers for meaningful analysis, categories may be combined where this is conceptually appropriate and can be justified in advance of analysis. For example, if numbers within individual ethnic minority groups are too small to support separate analyses, a broader comparison between participants from White backgrounds and those from minority ethnic backgrounds may be considered. Any such aggregation will be reported transparently, and descriptive statistics will, where possible, continue to be presented using the more disaggregated categories.

Where subgroup analyses are undertaken, findings will be reported with confidence intervals and interpreted cautiously, particularly where subgroup sizes are small. We will assess the presence of heterogeneous treatment effects in line with race equity, equality, diversity and inclusion considerations.

There is limited evidence about the relative effectiveness of mentoring programmes for those from ethnic minority backgrounds. As such, we will explore whether it might be possible to

evaluate whether SHIFT was equally effective for those from ethnic minority backgrounds compared to those from White backgrounds. This would likely be an underpowered analysis so caution should be applied when interpreting the results.

We will conduct this ANCOVA 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:

$$Y_i = \beta_0 + \beta_1 \text{Baseline}_i + \beta_2 \text{Group}_i + \delta_j \text{Ethnicity}_{ij} + \theta_j \text{Ethnicity}_{ij} \times \text{Group}_i + \varepsilon_i \varepsilon_i \sim N(0, \sigma^2)$$

For $i = 1, \dots, n$ young people, with $j = 1, \dots, J$ ethnicities

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 heterogeneous treatment effects by ethnicity. The total treatment effect for young people 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 SHIFT is differentially impactful for different ethnicities, we will also consider re-running the model in Section 6.2 for each sub-group separately where sample sizes permit (i.e. for young people from White, Asian, Black, Mixed and Multiple ethnic backgrounds, and Other ethnic backgrounds). If the sub-group specific treatment effects are similar, this will strengthen the findings from this exploratory analysis.

We will also explore the feasibility of evaluating whether the impact of SHIFT differs based on other characteristics including the presence of SEND, FSM eligibility, and care status. Subject to data availability, care status should distinguish between child looked after (CLA), child protection plan (CPP), child in need (CIN), none recorded and unknown. Because statuses can overlap over time, we will define a reference date and a mutually exclusive hierarchy (for example CLA, then CPP, then CIN) before analysis. If cells are small, a pre-specified any-social-care-status versus none comparison may be used.

6.5 Further analyses

We will conduct the following exploratory analyses.

The impact of positive relationships

The SHIFT theory of change identifies the development of a positive relationship with a trusted adult as a potential mechanism through which the intervention may influence delinquent behaviour. We will therefore conduct exploratory analysis to examine whether relationship quality is associated with the primary outcome and to inform understanding of

this aspect of the theory of change. Relationship quality will be measured at T2 using the Social Support and Rejection Scale (SSRS).

About the Social Support and Rejection Scale (SSRS)

The SSRS has four 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.

The SSRS is administered to all young people in the SHIFT group in relation to their SHIFT mentor. In the control group, young people are first asked whether they have a significant non-parental adult whom they see on a regular basis. Those who answer yes complete the SSRS in relation to that adult, while those who answer no do not complete the remaining SSRS items. We will therefore report the numbers and proportions in each trial arm with available SSRS data and describe the characteristics of those included in the SSRS analysis.

We will examine the association between each SSRS subscale and the T2 SRDS volume score, controlling for treatment allocation and baseline SRDS volume score. We will also consider a combined model including the SSRS dimensions simultaneously to explore which aspects of relationship quality appear most strongly associated with the primary outcome.

These analyses will be exploratory and descriptive rather than a formal mediation analysis. Because relationship quality and the primary outcome are both measured at T2, the temporal ordering of any association cannot be established. In addition, the control-group SSRS sample is necessarily selective because only young people who already have an ongoing relationship with a significant non-parental adult complete the measure. Findings will therefore be interpreted cautiously and used to inform refinement of the SHIFT theory of change rather than as estimates of a causal mechanism.

Dosage and fidelity

Fidelity to the model will be measured by the number of mentoring sessions participants receive, as well as the number of weeks over which they receive support. This will be compared to SHIFT's theory of change and its intended delivery model outlined within its manual, which aimed for a minimum of 11 sessions to be delivered within a six-month window.

We will also descriptively examine patterns of engagement across the delivery period using session-level attendance dates, if activity data is of sufficient quality. This will allow us to distinguish between young people who receive a similar overall number of sessions but experience different patterns of engagement, for example relatively consistent attendance, early disengagement, or a period of non-attendance followed by re-engagement. We will

report relevant descriptive indicators of continuity and gaps in engagement and consider these alongside total dosage when interpreting fidelity and implementation findings.

Analysis of dosage and fidelity will be descriptive and exploratory in nature, focusing on the number and timing of sessions received and patterns of engagement. These findings will inform the implementation and process evaluation and be considered alongside the CACE analysis described in Section 6.8, which will examine treatment effects among participants who complied with the intervention.

The impact on police contacts.

We will also explore the possibility of analysing reduced offending as measured by police data. We anticipate requesting data on police contacts, including type and timing of events (e.g., police contact, arrest, recorded offences, and convictions, all with associated dates). However, final data availability will be determined based on discussions with data providers.

If we can access robust and reliable police data, we will explore whether SHIFT had an impact on reducing contacts with police over and above that reported in the control group. We will do this by examining (1) whether a young person had any police-recorded contact/offence during the 26 weeks before and after randomisation, (2) the number of police-recorded contacts/offences during the 26 weeks before and after randomisation.

The analytical approach will reflect the distribution and structure of each outcome. Binary outcomes, such as whether a young person has any police-recorded offence/contact, will be analysed using logistic regression, with treatment effects reported as odds ratios with 95% confidence intervals. Count outcomes, such as the number of recorded offences/contacts, will be analysed using an appropriate count regression model such as negative binomial or Poisson; negative binomial regression will be preferred where the data are over-dispersed, with treatment effects reported as incidence rate ratios or equivalent exponentiated coefficients with 95% confidence intervals.

All police-data analyses will be exploratory. The final specification will depend on the completeness, coverage and consistency of the linked police data and the number of observed events. Any departures from the pre-specified approach necessitated by the structure or quality of the data will be documented and justified.

6.6 Imbalance at baseline

If randomisation has been successful, both treatment and control groups should be equivalent at baseline. As such, any imbalance will have occurred by chance. To check for and monitor imbalance, we will produce a table of descriptive characteristics for all young people who have completed a baseline questionnaire. We will also produce an equivalent table for

those who have completed a T2 questionnaire to check whether any attrition experienced throughout the trial may have introduced an imbalance.

These descriptive characteristics will include age, sex, ethnicity (collected by AudioActive colleagues) and the relevant primary and secondary outcomes collected through questionnaires. Categorical data will be summarised by numbers and percentages. Continuous data will be summarised by mean (SD) if data are normal and median [IQR] if data are skewed.

Depending on the quality and completeness of data received, we may also include other data collected by the AudioActive team including SEND data, care status, and FSM eligibility.

We will present a cross-tabulation of counts and percentages for each of the categories above against allocation group. For continuous variables, we will present the means and standard deviation. We will discuss any differences and their implications in the final report.

6.7 Missing data

Missing data may arise due to either item non-completion or sample attrition (i.e. young people who do not complete all items of the baseline or the T2 questionnaires). We will assess both the extent of missingness and patterns of missingness in the data in each arm of the trial. In line with YEF guidance (2021a) 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 (2021a). 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. We will not rely on the percentage threshold where there is substantive evidence that a small amount of missingness is informative (for example, loss to follow-up associated with custody or other circumstances).

- If outcome data is MAR conditional on co-variables, we will include these co-variables 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, with standard errors and 95% confidence intervals, will be reported in addition to estimates from the primary outcome analysis. 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 to assess the robustness of the estimated treatment effect to plausible departures from the assumptions made in the main analysis. The sensitivity analysis will examine how the estimated treatment effect changes under alternative assumptions about the outcomes of participants with missing follow-up data, and results will be compared with those from the primary complete-case and/or multiple-imputation analyses.

The assumptions used will be reported transparently, and conclusions will take account of whether the direction, magnitude or statistical significance of the treatment effect is materially altered under plausible MNAR scenarios.

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.

There is no universally agreed approach to analyses in the event of item non-completion. If a high proportion of cases would be excluded due to low rates of item noncompletion (for example, if most young people 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, in line with the YEF analysis guidance (2021a). We will also follow the YEF's updated guidance on missing data in the SRDS, when published.

6.8 Compliance

The primary analysis will be intention-to-treat: participants will be analysed according to randomised allocation irrespective of intervention receipt. For the purposes of the exploratory compliance analysis, a participant allocated to SHIFT will initially be defined as a complier if they receive at least 11 mentoring sessions within six months. This threshold was selected with reference to the intended delivery model. The realised distribution of session attendance will be described separately in the dosage and fidelity analysis in Section 6.5.

However, intent-to-treat analysis can underestimate the efficacy of the intervention if some young people 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 complied with the intervention, compared to those who would have complied had they been allocated to the intervention arm. 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 binary compliance (at least 11 sessions within six months versus fewer than 11) using randomised allocation as the instrumental variable, with baseline SRDS included for precision. The strength of the instrument will be assessed using the appropriate first-stage statistic. Continuous session dosage will be examined separately in the exploratory dosage analysis in Section 6.5 and will not be used as the binary CACE compliance definition.
2. The predicted values for compliance generated in the first stage will be used in the outcome model to estimate the CACE effect on the T2 SRDS volume score, adjusting for baseline SRDS volume score.

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 effect estimates with its standard error and 95% confidence interval will be provided in the final report.

6.9 Intra-cluster correlation

This is not a clustered randomised controlled trial. As such, we will not be calculating intra-class correlations.

6.10 Presentation of outcomes

The 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, as specified in the primary outcomes model.

- sd_{pooled} is the unconditional pooled standard deviation of the two groups.⁵

Hedges' g will be used as the standardised effect-size measure. It is closely related to Cohen's d but applies a correction for small-sample bias, with the difference between the two becoming smaller as sample size increases. Hedges' g standardises the adjusted mean difference using the pooled standard deviation from the treatment and control groups. If there is substantial heterogeneity in outcome variances between the two arms, the pooled standard deviation may be less representative. In this circumstance, Glass' delta, which standardises the adjusted mean difference using the control-group standard deviation only, may also be reported as a sensitivity effect size (Lipsey & Wilson, 2001). Any use of Glass' delta will be justified and reported alongside the primary Hedges' g estimate rather than selected on the basis of statistical significance.

We will report the statistical uncertainty associated with the effect sizes using both 95% confidence intervals (CIs) and two-tailed p -values, with statistical significance assessed at the conventional threshold of $p < 0.05$. 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.
- g_{se} is the standard error of the g statistic (noted as ES above).

For the primary and secondary outcomes we will report the adjusted mean difference between trial arms and the standardised effect size (Hedges' g), each with 95% confidence intervals, together with two-sided p -values where appropriate. We will interpret point estimates alongside their uncertainty and the existing evidence base. Findings will also be triangulated with the implementation and process evaluation evidence on delivery quality and context, theoretical mechanisms, and the experiences of young people, staff and referrers who participate in semi-structured interviews.

⁵ $sd_{pooled} = \sqrt{\frac{(n_1-1)sd_1^2 + (n_2-1)sd_2^2}{n_1 + n_2 - 2}}$, 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.

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.
- Dong, N. and Maynard, R. A. (2013). *PowerUp!:* A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, 6(1), 24-67. doi: 10.1080/19345747.2012.673143
- Fonagy P, Butler S, Cottrell D, Scott S, Pilling S, Eisler I, Fuggle P, Kraam A, Byford S, Wason J, Ellison R, Simes E, Ganguli P, Allison E, Goodyer IM. (2018). Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): a pragmatic, randomised controlled, superiority trial. *Lancet Psychiatry* 5 (2), 119-133, doi: 10.1016/S2215-0366(18)30001-4.
- Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Child and Adolescent Psychiatry*, 40(11), 1337-1345.
- Goodman, R., 2005. *The Strengths and Difficulties Questionnaire*. Available at: <https://www.sdqinfo.org/a0.html> [Accessed 17/8/23].
- Hayes, B., 2011. An adventure in the Nth dimension. *American Scientist*, 99(6), pp.442-446.
- Humayun, S., Herlitz, L., Chesnokov, M., Doolna, M., Landau, S., Scott S. (2017). Randomised controlled trial of Functional Family Therapy for offending and antisocial behaviour in UK Youth. *Journal of Child Psychology and Psychiatry* 58(9), 1023-1032.
- Humayun et al. (2019). *Unpublished trial data analysis shared with the evaluation team for the purposes of informing power calculation assumptions*.
- 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.
- Jolliffe, D. and Farrington, D.P. (2008). The influence of Mentoring on Reoffending *Swedish National Council for Crime Prevention*. Available at:
- Karcher, M.J. and Nakkula, M.J., 2010. Youth mentoring with a balanced focus, shared purposes, and collaborative interactions. *New Directions for Youth Development* 2010(126), pp.1-32.
- Lipsey, M. and Wilson, D. (2001) *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McAra, L., McVie, S. (2005). The usual suspects? Street-life, young people and the police. *Criminal Justice* 5(1), 5-36.

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.

Smith, D.J and McVie, S. (2003). Theory and method in the Edinburgh study of youth transitions and crime. *British Journal of Criminology* 43, 169–95.

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, (2021a). *Analysis Guidance*. Available at: <https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf> [Accessed 20 /10/23].

Youth Endowment Fund, (2021b). *Core measurement guidance: Self-Report Delinquency Scale*. Available at: <https://res.cloudinary.com/yef/images/v1623145465/cdn/19.-YEF-SRDS-guidance/19.-YEF-SRDS-guidance.pdf>. [Accessed 20 /10/23].

Youth Endowment Fund. (2022a) *Core Measurement Guidance: Strengths and Difficulties Questionnaire* <https://youthendowmentfund.org.uk/wp-content/uploads/2022/04/18.-YEF-SDQ-guidance-April-2022.pdf> Accessed 15/09/23.

Youth Endowment Fund. (2022b). *Outcomes Framework* <https://youthendowmentfund.org.uk/wp-content/uploads/2022/08/YEF-Outcomes-Framework-August-2022.pdf> Accessed 15/09/23.



youthendowmentfund.org.uk



hello@youthendowmentfund.org.uk



[@YouthEndowFund](https://twitter.com/YouthEndowFund)

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413
