STATISTICAL ANALYSIS PLAN

Spark2Life Meaningful Mentoring Programme: a randomised controlled trial efficacy study with an internal pilot

ICF Consulting Services Ltd

Principal investigator: Dr Matt Barnard

YOUTH ENDOWMENT FUND

Spark2Life's meaningful mentoring programme to reduce the risk of violence and offending: a two-armed randomised controlled trial with an internal pilot

Statistical analysis plan

Evaluating institution: ICF Consulting Services Ltd
Principal investigator(s): Dr Matt Barnard

YEF statistical analysis plan

Project title	Spark2Life Meaningful Mentoring Programme: a randomised controlled trial efficacy study with an internal pilot
Developer (Institution)	Spark2Life
Evaluator (Institution)	ICF Consulting Services Ltd
Principal investigator(s)	Dr Matt Barnard
SAP author(s)	Robert Wishart, Dr Andre Butler, Dr Matt Barnard, Dr Lucy Loveless
Trial design	Two-armed multi-site randomised controlled trial with random allocation at the individual level
Trial type	Efficacy trial with an internal pilot and an implementation and process evaluation
Evaluation setting	Community

Target group	Children and Young People (CYP) aged 11 to 18 years old, who are involved in youth violence, gang activity, and/or crime as a perpetrator or victim
Number of participants	704
Primary outcome and data source	Self-reported offending (measured by the volume score on the Self-Report Delinquency Scale)
Secondary outcome and data source	Behavioural and emotional difficulties (measured by the total difficulties score on the Strengths and Difficulties Questionnaire) Prosocial behaviour (measured by the prosocial subscale of the Strengths and Difficulties questionnaire) Presence and quality of relationship with an adult who isn't a parent (measured by the Social Support and Rejection Scale)

SAP version history

Version	Date	Changes made and reason for revision
1.0 [*original*]		N/A

Changes from the protocol

The evaluation's protocol (Barnard, et al., 2024) outlined the decisions made at the inception of the trial for the trial design. Between the protocol's publication and the drafting of this Statistical Analysis Plan (SAP), there were a few revisions to the trial design, agreed with the YEF.

The primary change to the trial design is an extension to the recruitment period, and a reduction in the recruitment target. This was introduced due to challenges in achieving a sufficient number of referrals and recruitment within the original trial period. This has some

implications for the power and sample size section, discussed in greater detail later in this document.

In addition to changes to the recruitment timeline, the referral partners were also expanded. Whereas the protocol outlined an intention to recruit through Youth Offending Service (YOS) teams, the recruitment has now been expanded to include other settings, including mainstream schools, pupil referral units, Local Authority social care and early help services. We refer to these as “settings” collectively throughout the SAP.

Finally, as outlined in the protocol, preliminary analysis has been conducted to refine the choice of analysis models for some outcomes. This is discussed further in the Analysis section of this document.

Table of contents

SAP version history 2

Table of contents 3

Introduction 4

Design overview 4

Sample size calculations overview 5

Analysis 8

References 16

Appendix 1: Round 1 analysis 17

Introduction

Meaningful Mentoring is a community-based, one-to-one mentoring intervention delivered by Spark2Life to children and young people (CYP) aged 10–18 who are involved in, or at risk of, youth violence and offending. The programme provides up to twelve months of sustained support through regular, typically weekly, face-to-face sessions between the CYP and a trained mentor. Delivery is intentionally flexible and non-manualised, enabling mentors (many of whom have lived experience of the issues faced by mentees) to tailor support to the individual needs, circumstances and preferences of each young person, while maintaining a consistent emphasis on building a trusting, supportive and empathetic relationship. The intervention takes a competent, person-centred and trauma-informed approach. Mentoring is highly relational and client-led, with mentors adapting their approach to reflect the lived experiences and needs of CYP, many of whom have experienced instability, adverse experiences and disengagement from services. The programme aims to address a number of risk and protective factors including relationships, family dynamics, peer influences and sense of purpose. It aims to strengthen agency, improve socio-emotional skills, and support more positive decision-making ultimately leading to improved behavioural outcomes over the longer term. A more detailed description of the intervention can be found in the evaluation protocol (Barnard, et al., 2024).

This study aims to evaluate Meaningful Mentoring through a two-armed randomised controlled trial (RCT), randomised at the individual level and with the control group receiving business-as-usual from youth offending services, children's social care services or alternative provision (AP) schools, including pupil referral units (PRUs) or a combination of all three. The study incorporates an implementation and process evaluation, which involves collecting quantitative and qualitative data from CYP in both intervention and control groups and from mentors and management staff. In addition, the study will include a cost-consequence analysis. The rationale behind the approach is that RCTs are an effective way of assessing the net impact of an intervention and a two-armed, individually randomized approach was selected as the most feasible and efficient way of generating a sufficient sample size given the capacity of Spark2Life.

Design overview

Trial design, including number of arms	Two-arm, multi-site randomised controlled trial
Unit of randomisation	Individual CYP

Stratification variables (if applicable)		Setting (Youth Offending Service Team; mainstream schools, pupil referral units, Local Authority social care and early help services)
Primary outcome	variable	Self-reported offending
	measure (instrument, scale, source)	Self-Report Delinquency Scale (volume of offending score, twelve months post-randomisation)
Secondary outcome(s)	variable(s)	Behavioural and emotional difficulties Prosocial behaviour Presence and quality of relationship with a trusted adult who is not a parent
	measure(s) (instrument, scale, source)	Strengths and Difficulties Questionnaire – total difficulties score, twelve months post-randomisation Strengths and Difficulties Questionnaire – prosocial behaviour score, twelve months post-randomisation Social Support and Rejection Scale, twelve months post-randomisation
Baseline for primary outcome	**variable**	Self-reported offending
	measure (instrument, scale, source)	Self-Report Delinquency Scale (volume of offending score)
Baseline for secondary outcome	**variable**	Behavioural and emotional difficulties Prosocial behaviour Presence and quality of relationship with a trusted adult who is not a parent
	measure (instrument, scale, source)	Strengths and Difficulties Questionnaire – total difficulties scale Strengths and Difficulties Questionnaire – prosocial behaviour score Social Support and Rejection Scale

Sample size calculations overview

The sample size for the trial is based on achieving a minimum detectable effect size (MDES) of 0.20. The literature indicates that the mean effect size of mentoring on reoffending is 0.212 (Gaffney, et al., 2022) and the YEF recommends studies are powered to achieve an MDES of 0.20, which aligns well with the existing research literature and hence has been selected for this evaluation.

Power calculations were conducted using PowerUp! (Dong & Maynard, 2013) for a two-level blocked-individual random assignment design using fixed effects (Sheet 2.2 BIRA2_1f). The power calculations were based on the following assumptions:

- A type-one (false positive) error rate of 0.05.
- A type-two (false negative) error rate of 0.20 (synonymous with power of 0.80).
- Two-tailed statistical significance testing.
- A 1:1 allocation ratio between intervention and control.
- The variance in the outcome at endline that is expected to be explained by the baseline measure and blocking covariates is 0.20. We estimate this to be roughly equivalent to a pre-post correlation of 0.45. This is a reasonable assumption given that the same outcome measure is being used at baseline and endline, twelve months apart and is similar to the pre-post correlation assumed by another YEF trial (Mentoring Multi-Site Trial). Based on round 1 data, the pseudo-R^2 was 0.38, yielding an estimated pre-post correlation of 0.60. The randomisation power calculations are updated accordingly.

To reach an MDES of 0.20, a recruitment target of 704 CYP (352 intervention: 352 control) was agreed for the trial at protocol stage. Since the protocol was published, a revised recruitment target and extended recruitment timeline was agreed between YEF and Spark2Life due to slower than anticipated recruitment. The revised recruitment target was 600 CYP (300 intervention; 300 control) with recruitment extended to end of November 2026. Additionally, recruitment was expanded to include other referral partners. The protocol outlined that recruitment would be through Youth Offending Service Teams. Whilst these remain referral partners, other settings (mainstream schools, pupil referral units, Local Authority social care and early help teams) were also added to ensure a sufficient number of CYP could be referred to the trial. Therefore, randomisation is blocked by setting, rather than Youth Offending Service team alone. The power calculations presented as "Extended recruitment" are therefore aligned to this revised recruitment target, rather than those in the published protocol. The protocol power calculations were adjusted the sample size to account

for attrition between baseline and endline of 10%. However, following YEF guidance, the power calculations presented here are based on the recruited sample without attrition.

At the time of writing the SAP, the recruitment is ongoing, and therefore we use the projected sample size numbers.

The sample size calculations are summarized in Table 1 below. Note that the power calculations in Table 1 are presented as a Hedge's g effect sizes, but the primary analysis model will be presented as an incidence rate ratio (IRR). Therefore, we approximate the equivalent IRR as $g = \ln(IRR) \times \frac{\sqrt{3}}{\pi}$. At protocol stage, this is an IRR of 0.70, equivalent to a 30% reduction in the number of offences. At randomization, we approximate this as an IRR of 0.72, equivalent to a 28% reduction in the number of offences.

Table 1 Power calculations

		Protocol	Extended recruitment	Randomisation
Minimum Detectable Effect Size (MDES)[1]		0.19 (0.20) [2]	0.20	0.18
Pre-test/ post-test correlations[3]	level 1 (participant)	0.45	0.45	0.60
	level 2 (cluster)	n/a	n/a	n/a

[1] This is presented as a *Hedge's g* effect size, but the primary analysis model will be presented as an incidence rate ratio (IRR). Therefore, we approximate the equivalent IRR as $g = \ln(IRR) \times \frac{\sqrt{3}}{\pi}$. At protocol stage, this is an IRR of 0.70, equivalent to a 30% reduction in the number of offences. At randomization, we approximate this as an IRR of 0.72, equivalent to a 28% reduction in the number of offences.

[2] At protocol, the sample size was presented to include expected attrition of 10%. However, YEF analysis guidance states that the sample size calculations should be based on the recruited rather than achieved sample size. Based on the recruited sample size expected at protocol stage, of 704, the MDES would have been smaller, which we present here for consistency, with the 10% attrition values presented in brackets.

[3] Whilst the YEF template requests the pre- post- test correlation, in practice it is the variance explained (R^2) that influences the power. There is no formula to transform a correlation to an R^2 as they represent different concepts. However, we use the square-root of our assumed R^2 (0.20 at protocol, 0.38 at SAP) to obtain an approximation of the expected pre-post correlation.

		Protocol	Extended recruitment	Randomisation
Intracluster correlations (ICCs)	level 1 (participant)	n/a	n/a	n/a
	level 3 (cluster)	n/a	n/a	n/a
Alpha		0.05	0.05	0.05
Power		0.80	0.80	0.80
One-sided or two-sided?		Two-sided	Two-sided	Two-sided
Average cluster size		n/a	n/a	n/a
Number of clusters	Intervention	n/a	n/a	n/a
	Control	n/a	n/a	n/a
	Total	n/a	n/a	n/a
Number of participants	Intervention	352 (317)	300	300
	Control	352 (317)	300	300
	Total	704 (634)	600	600

Analysis

All analyses will be conducted in the latest available version of Stata at the time of analysis. At the time of writing the SAP, this is Stata 19 MP. Analysis will be conducted in March 2028.

Primary outcome analysis

The primary outcome is self-reported offending, measured by the Self-Reported Delinquency Scale Volume of Offending Score (SRDS), measured twelve months post-randomisation. Respondents are asked a series of questions about the number of incidents that have occurred in the last twelve months, separately for different types of offence. The volume score is based on the total number of incidents across all offence types that occur over the last twelve months and follows the scoring outlined in the YEF guidance on the use of the SRDS (Youth Endowment Fund, 2021). Item 16 is not included in the generation of the volume score. Following YEF guidance, unadjusted means with 95% confidence intervals and underlying distributions will be presented at baseline and endline in the final evaluation report. In addition to this, the impact will be presented as an effect size with 95% confidence intervals and p-values.

The primary analysis will be conducted on an intention-to-treat basis and will use a linear regression model, where self-reported delinquency twelve months after randomisation is the dependent variable, regressed against self-reported delinquency at baseline, intervention allocation (which is interpreted as the impact estimate) and fixed effects for the setting within which they were randomised.

At protocol, we proposed a linear regression model as this is consistent with other YEF evaluations using the same outcome measure. However, in the general population, self-reported delinquency follows a Poisson distribution. As outlined in the protocol, at the end of the internal pilot, we assessed the outcome distribution to determine whether analysis using a Poisson regression would be more appropriate. We will re-assess our assumptions at analysis stage to ensure a Poisson regression is appropriate (including overdispersion).

This analysis is presented in Appendix 1 and it indicates that a Poisson regression would be most appropriate for the primary analysis. The primary analysis model will have the following specification.

$$SRDS_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 SRDS_{ij,t0} + \beta_3 X_j + \varepsilon_i$$

Where the SRDS at baseline (t0) and follow-up (t1) is measured by the Self-Report Delinquency Scale (volume measure), for individual i in setting j. Allocation is a binary indicator of intervention or control group allocation, where the coefficient β1 is interpreted as the treatment effect. Xj represents the vector of settings. Alpha represents the intercept and epsilon the random error. The regression will be estimated using Huber-White robust standard errors. As there will be a large number of settings, the analysis will be conducted using the ***ppmlhdfe*** command in Stata, with the following syntax:

```
ppmlhdfe f_vds i.allocation b_vds, absorb(setting) vce(robust) irr
```

As the primary outcome is a count variable, the effect size will be presented as an incidence rate ratio (IRR). Further detail on the effect size calculations is outlined later in the **Presentation of Outcomes** section.

Sensitivity analysis

Sensitivity analysis will be conducted in line with YEF guidance. Specifically, we will analyse a "saturated model", adding additional covariates. The full list of independent variables is:

- Baseline volume of delinquency score
- Random allocation
- Setting
- Sex
- Major ethnicity group (White, Asian, Black, Mixed, Other)
- Any physical or mental health conditions or illnesses lasting or expecting to last more than twelve months (binary)
- Baseline Strengths and Difficulties Questionnaire (SDQ) total difficulties score
- Baseline Social Support and Rejection Scale (SSRS) total score

A further sensitivity analysis will be implemented, re-estimating the primary analysis but including the mentor-level clustering in the intervention group. Specifically, whilst individuals are blocked by settings, the intervention group will also experience clustering through the mentors. Mentors will have multiple mentees, and these are not necessarily all within the same site. This means we have a partially cross-classified clustered design and not accounting for the partial cross-classification could lead to an underestimate of standard errors. The sensitivity analysis will use a multilevel Poisson model, with the following specification.

$$SRDS_{ijk} = \beta_0 + \beta_1 Allocation_{ijk} + \beta_2 SRDS_{jk} + u_{jk} + w_k + e_{ijk}$$

Where u_{jk} is the random intercept for mentors and w_k the random intercept for sites. This would be estimated in Stata as follows[4]:

```
mepoisson f_vds i.allocation b_vds || _all || r.mentor: || setting:, irr
```

Secondary outcome analysis

Secondary outcome analysis will be conducted to assess the impacts using the Strengths and Difficulties Questionnaire total difficulties score and prosocial subscale; and the Social

[4] Each individual in the control group would have a separate, unique, mentor ID assigned to properly account for partial cross-classification.

Support and Rejection Scale. There is no adjustment to the secondary outcome analysis to account for multiple hypothesis testing. The secondary outcome analysis should be considered exploratory, and we establish the hierarchy of secondary outcomes below.

1. Total difficulties score
2. Prosocial subscale
3. Social Support and Rejection Scale

The SDQ measures emotional and behavioural difficulties and has been widely used in clinical and research settings. It has good psychometric properties and is comprised of four subscales relating to emotional and behavioural issues which are; emotional symptoms, conduct problems, hyperactivity/inattention and peer problems. The total difficulties score has a scale of 0-40, and is the sum of each of the SDQ subscales. The prosocial behaviour subscale has a range of 0-10. The Social Support and Rejection Scale is recommended by the National Mentoring Resource Center and is also being used in the YEF-funded evaluation of the STEER mentoring programme. The scale consists of 22 items, each on a five-point scale, assessing four dimensions of social support and social rejection that characterise the relationship between young people and a trusted adult such as a mentor. The SSRS has a range of 0-88,

Following the primary analysis, unadjusted means with 95% confidence intervals, underlying outcome distributions and the effect size with 95% confidence intervals and p-values will be reported. The two secondary outcomes from the SDQ will be analysed using a linear regression model as follows:

$$Total\ difficulties_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 Total\ difficulties_{ij,t0} + \beta_3 X_j + \varepsilon_i$$

$$Prosocial_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 Prosocial_{ij,t0} + \beta_3 X_j + \varepsilon_i$$

In these models, the coefficient β_1 is interpreted as the impact estimate. All other terms are consistent with the primary analysis model, substituting the dependent variable and the baseline measure of the outcome with the relevant secondary outcome.

The regression models will use Huber-White robust standard errors. Due to the large number of fixed-effects the regression will be estimated using the ***reghdfe*** command in Stata.

```
reghdfe f_td i.allocation b_td, absorb(setting) vce(robust)

reghdfe f_ps i.allocation b_ps, absorb(setting) vce(robust)
```

The initial analysis from the first round of recruitment (see Appendix 1) indicated that the SSRS is also over dispersed and will therefore be analysed using a Poisson regression[5], with the following specification.

$$SSRS_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 SSRS_{ij,t0} + \beta_3 X_j + \varepsilon_i$$

As there will be a large number of settings, the analysis will be conducted using the ***ppmlhdfe*** command in Stata, with the following syntax:

```
ppmlhdfe f_ssrs i.allocation b_ssrs, absorb(setting) vce(robust) irr
```

Subgroup analyses

Subgroup analyses will explore whether impacts on the primary outcome vary by disability status[6] and by ethnicity.

A dichotomous analysis of White CYP against CYP from minoritised ethnic groups would be over-simplistic and hide nuances in the data. Based on the target population of beneficiaries, we expect that a large proportion of individuals recruited to the trial will be from Black, Asian or other minority ethnicity backgrounds. Whilst we intended to separately analyse the outcome for different ethnicity groups tested against a consistent reference group (White CYP) Table 2 below outlines the proportions of CYP by ethnicity group based on data collected in the internal pilot (round 1 of recruitment).

Table 2 Proportion of round 1 participants by ethnicity group

Ethnic group	Count	Percentage (%)
Asian or Asian British	†	†
Black, Black British, Caribbean or African	75	43
Mixed or multiple ethnic groups	34	19

[5] Further detail on this decision is outlined in the Appendix. We opt for Poisson over negative binomial as Poisson is still consistent with overdispersion, and negative binomial models rely on stricter assumptions that are often violated (Wooldridge, 1997).

[6] In the protocol, this was indicated as Special Educational Need (SEN). However, the questionnaire wording asks about chronic health conditions or disabilities rather than sepcifc Special Educational Needs. Therefore disability status feels a more appropriate label.

White	48	27
Other Ethnic Group	†	†
Total	176	100

† indicates that the cell has been supressed to avoid statistical disclosure. This means one or more of the suppressed cells has a count of less than ten individuals. Additional cells may be supressed to ensure cell counts of less than ten cannot be back calculated.

Using this as a guide for the likely sample sizes at the end of the analysis we anticipate that the number of individuals in the Asian or Asian British, and Other Ethnic Group categories would not be sufficient to conduct a robust subgroup analysis due to statistical power. Even with collapsed categories, these analyses are underpowered and should therefore be interpreted with caution (as are the analyses by disability). Therefore, the subgroup analysis will be separated into the following categories: White (reference group); Black, Black British, Caribbean or African; All other ethnic groups (Asian or Asian British; Mixed or multiple ethnic group; and Other Ethnic Group will be combined into a single category). Even after combining categories, the subgroup analyses are likely to have lower statistical power which may make it harder to detect statistically significant differences in the relative effectiveness of the intervention for different ethnicity groups. For each sub-group (ethnicity subgroups, disability), the unadjusted means and 95% confidence intervals will be presented along with the effect size, 95% confidence intervals and p-values.

The analysis by disability status is based on whether the individual reports having any physical or mental health conditions or illnesses lasting or expected to last more than twelve months at baseline. Approximately 16% of CYP with non-missing data (n = 161) of the round one trial participants had a physical or mental health condition or illness. Whilst a follow-up question about the impact of the condition or illness on day-to-day life (not at all, a little, a lot) was included in the survey questionnaire, the cell counts are too small for robust subgroup analysis.

Given that both sets of subgroup analysis are expected to be underpowered, the subgroup analysis is exploratory. The subgroup analyses will be conducted using an interaction model, with the following regression specifications.

$$SRDS_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 Ethnicity_{ij} + \beta_3 Allocation * Ethnicity_{ij} + \beta_4 SRDS_{ij,t0} + \beta_4 X_j + \varepsilon_i$$

$$SRDS_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 Disability_{ij} + \beta_3 Allocation * Disability_{ij} + \beta_4 SRDS_{ij,t0} + \beta_4 X_j + \varepsilon_i$$

For each analysis, the vector of coefficients β_3 reflects the interaction between random allocation and each othe subgroups, whilst β_1 represents the impact effect for the reference group (white, no disability).This will be analysed in Stata using the following syntax:

```
ppmlhdfe f_vds i.allocation i.ethnicity i.allocation#i.ethnicity
b_vds, absorb(setting) vce(robust) irr
```

```
ppmlhdfe f_vds i.allocation i.disability i.allocation#i.disability
b_vds, absorb(setting) vce(robust) irr
```

Further analyses

The primary outcome will also be analysed using a mediation analysis, which will focus on the mentor relationship, as measured by the SSRS. We will do this by exploring how the effect of the intervention on offending is mediated by an increase in the presence and quality of a trusted adult. This will be estimated in Stata using the **mediate** command as follows:

```
mediate (f_vds b_vds, poisson) (f_ssrs_total b_ssrs_total)
(allocation)
```

In this syntax, f_vds is the volume of deliqeuncy score at follow-up, b_vds is the volume of delinquency score at baseline, f_ssrs is the SSRS score at follow-up and b_ssrs is the SSRS score at baseline. The models are specified separately for the poission and OLS components.

Additionally, we will assess treatment effect heterogeneity between sites. Specifically, this will be the site the CYP was in when they were randomised (for example, this could be a Youth Offending Service (YOS) team or a Pupil Referral Unit (PRU)). This will be assessed by interacting the trial arm with the fixed effects for the site. This will help us to understand what the variation in effectiveness is like, and may help further understand what conditions the intervention is more or less effective in. This will be reported in a way that does not disclose the identity of the sites.

$$SRDS_{ij,t1} = \propto + \beta_1 Allocation_{ij} + \beta_2 Site_{ij} + \beta_3 Allocation * Site_{ij} + \beta_4 SRDS_{ij,t0} + \varepsilon_i$$

The individual interaction effects will be presented in forest plot with 95% confidence intervals, sorted from the smallest to largest impact.

Interim analyses and stopping rules

Interim analysis was conducted at the end of the internal pilot to aid analytical decisions made in this Statistical Analysis Plan. The results of these analysis are detailed in Appendix 1.

Longitudinal follow-up analyses

No longitudinal follow-up analysis is planned.

Imbalance at baseline

Imbalance at baseline will be assessed descriptively. Specifically:

- Continuous outcomes will be reported as means and standard deviations.
- Binary and categorical outcomes will be reported as counts and percentages.
- Count outcomes will be reported as mean and variance.

The analyses will be presented in a table with their means/proportions/medians and differences by allocation as effect sizes.

The characteristics that will be tested for imbalance at baseline are outlined in Table 3.

Table 3 Characteristics assessed for imbalance at baseline

Characteristic	Variable type
Sex	Binary
Ethnicity (major ethnicity group)	Categorical
Disability	Binary
Baseline volume of delinquency scale	Count
Baseline total difficulties score	Continuous
Baseline prosocial behaviour score	Continuous
Baseline social support and rejection scale	Count

Missing data

If fewer than 5% of randomised CYP are excluded from the primary analysis model due to missing data, no missing data analysis will be conducted.. If 5% or more of randomised CYP are excluded from the primary analysis model due to missing data, we will assess if there are observable patterns in the "missingness" mechanism using a logistic regression model, where the dependent variable is a binary indicator of missing outcome data, regressed against all available covariate data. In this model, missing covariate data will be replaced with a "missing" category for categorical variables, the grand mean for continuous variables and median for count variables. If the logistic regression model indicates that there are observable patterns in the "missingness" mechanism, then we assume data is missing at random (MAR) as it is not possible to explicitly test for data missing not at random (MNAR). If the data is

assumed to be MAR, we will conduct multiple imputation through chained equations (MICE) to impute missing values as a sensitivity analysis (for example Azur, et al., 2011). If there are no observed associations, data is assumed to be missing completely at random (MCAR).

Imputed datasets will be created using the **mi** suite of commands in Stata. The model will include the primary outcome and all the available covariates and auxiliary variables. The first 100 imputations will not be used ("burn in") to ensure that the imputation has reached a stable distribution. Whilst a stable distribution might be achieved with fewer imputations, the benefits of a greater number of imputations are perceived to outweigh the increased computation time needed given the expected sample size. A total of 100 imputed datasets will be created to ensure the Monte Carlo error of the coefficient and standard error are reproducible within a reasonable limit (van Buuren, 2018).

Compliance

Compliance will not be assessed as part of the impact evaluation; a complier average causal effect will not be estimated. This is because the length of mentoring that is required is likely to vary from person to person, so compliance is not synonymous with "dosage". Given the flexible and tailored nature of the mentoring being provided, it is not possible to create a clear binary or continuous measure of compliance. Compliance will instead be assessed through the Implementation and Process evaluation using quantitative and qualitative analysis. This will include presenting information about number and frequency of mentoring sessions attended and the proportion who dropped out before mentoring was completed and before planned goals were met.

Presentation of outcomes

The outcomes, their corresponding type, analysis model and reporting metric is outlined in Table 4. Alongside the effect size, 95% confidence interval and p-value, the unadjusted means and 95% confidence intervals will also be presented.

Table 4 Outcomes, outcome type, analysis model and reporting metrics

Outcome	Type	Analysis model	Reporting metric
Self-Report Delinquency Scale (volume of offending score)	Count	Poisson	Incident rate ratio
Behavioural and emotional difficulties	Continuous	Linear regression	Hedge's g
Prosocial behaviour	Continuous	Linear regression	Hedge's g
Presence and quality of relationship with a	Count	Poisson	Incident rate ratio

trusted adult who is not a parent			

Note that, following YEF analysis guidance, alongside any effects presented as incident rate ratios, an equivalent Hedge's g effect size will be reported, using the formula below (Borenstein, et al., 2009).

$$g = \ln(IRR) \times \frac{\sqrt{3}}{\pi}$$

The count outcomes will be presented as *Incidence Risk Ratios (IRR)* with confidence intervals. The formula is as follows:

$$IRR = \frac{Incident\ rate_T}{Incident\ rate_C} = \frac{\frac{Number\ of\ events_T}{N_T}}{\frac{Number\ of\ events_C}{N_C}}$$

The continuous outcomes will be presented as *Hedge's g* effect sizes with confidence intervals using the formula from (Hedges, 2007) for the effect size d_t for designs with unequal sample sizes. Following YEF reporting guidance, the adjusted difference in means will be reported in an appendix. The effect size g_t is estimated as follows:

$$g_t = J \times \left(\frac{\overline{Y^T} - \overline{Y^C}}{S_t}\right)$$

Where $\overline{Y^T}$ and $\overline{Y^C}$ are the grand means of the intervention and control groups respectively. The correction factor J is given by:

$$J = 1 - \left(\frac{3}{4(N^T + N^C - 2) - 1}\right)$$

The pooled standard deviation, S_t is defined as:

$$S_T = \sqrt{\frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n_i^T}(Y_{ij}^T - \bar{Y}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n_i^C}(Y_{ij}^C - \bar{Y}^C)^2}{N-2}}$$

To calculate the confidence intervals, the variance term is estimated as:

$$V_{\{g_t\}} = \left(\frac{N_T + N_C}{N_T N_C}\right) + \frac{g_t^2}{2(N-2)}$$

The confidence intervals are calculated as follows:

$$g_t - zV_{\{g_t\}} \leq \delta T \leq g_t + zV_{\{g_t\}}$$

Where z is the critical value.

References

Azur, J. M., Stuart, E. A., Frangakis, C. & Leaf, P. J., 2011. Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods of Psychiatric Research,* 20(1), pp. 40-49.

Barnard, M., Loveless, L. & Wishart, R., 2024. *Evaluation protocol: Spark2Life's meaningful mentoring programme to reduce the risk of violence and offending: a two-armed randomised controlled trial with an internal pilot,* s.l.: Youth Endowment Fund.

Dong, N. & Maynard, R., 2013. PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness,* 6(1), pp. 24-67.

Gaffney, H., Darrick, J. & White, H., 2022. *Mentoring Programmes: YEF Technical Report,* s.l.: Youth Endowment Fund.

Hedges, L. V., 2007. Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioural Statistics,* 32(4), pp. 341-370.

Leckie, G. et al., 2020. Partitioning variation in multilevel models for count data. *Psychol Methods,* 25(6), pp. 787-801.

van Buuren, S., 2018. *Flexible Imputation of Missing Data.* 2 ed. New York: Chapman and Hall/CRC.

Youth Endowment Fund, 2021. *Core measurement guidance: Self-reported delinquency scale (SRDS),* s.l.: Youth Endowment Fund.

Appendix 1: Round 1 analysis

This appendix outlines analysis conducted using the first round of trial participants.

Primary outcome distribution

The primary outcome, SRDS, has been used in several YEF evaluations. However, these evaluations have used a linear regression model, whilst the outcome distribution in the general population is Poisson distributed. Therefore, we assessed whether a linear regression model was suitable for the primary analysis. Figure 1 and Figure 2 show the outcome distributions observed for the first round of trial participants at baseline and follow-up respectively.

Figure 1 Baseline volume of delinquency scale distribution

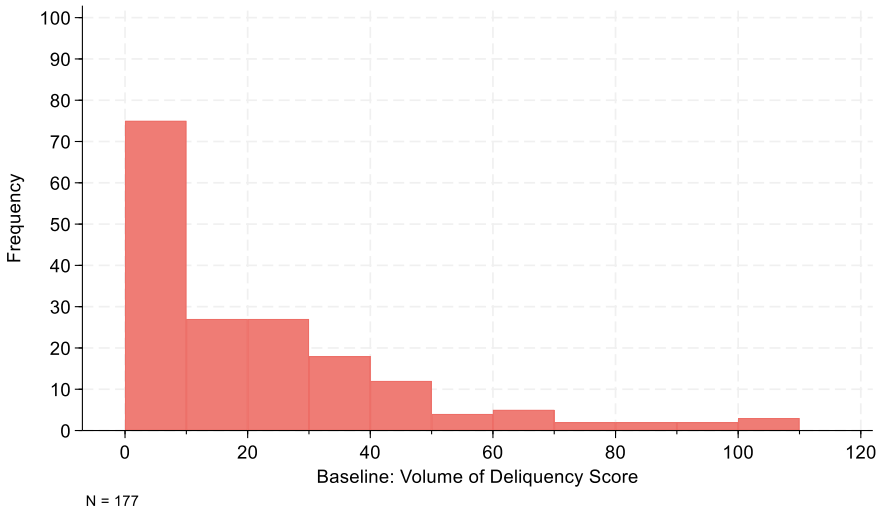
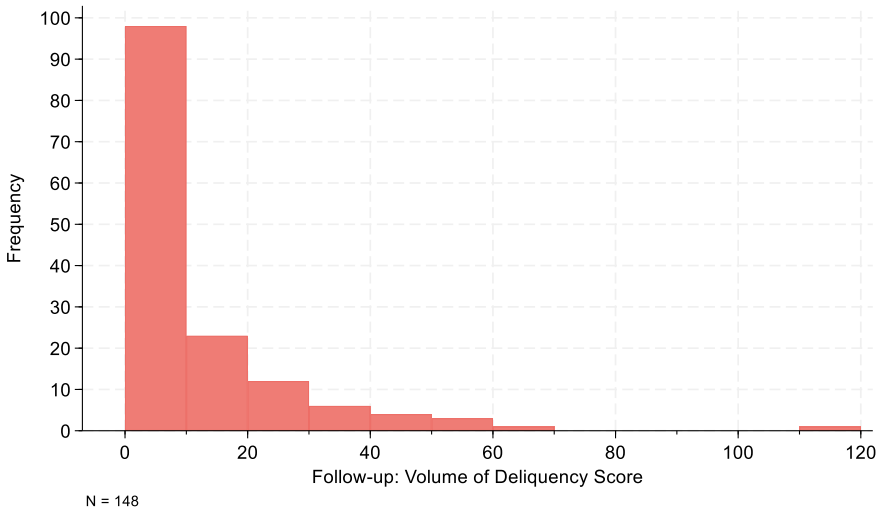
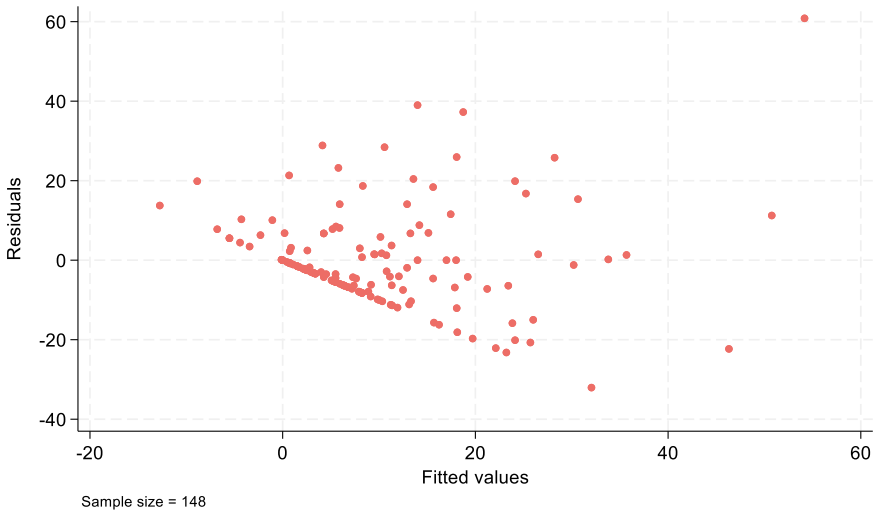


Figure 2 Follow-up volume of delinquency scale distribution



A large proportion of trial participants exhibit zero values; 14% at baseline and 45% at follow-up. These distributions confirm that for our sample, the outcome follows a Poisson distribution. In addition, we estimated a residual-versus-fitted plot from a linear regression model where the dependent variable was the volume of delinquency scale at follow-up, regressed against treatment allocation, baseline volume of delinquency scale and fixed effects for sites.

Figure 3 Residual-versus-fitted plot – Volume of delinquency scale



The distribution should be totally random if the assumptions of the linear regression model hold. The distribution is fanned, indicating that linear regression is not suitable.

Based on these analyses, the primary outcome analysis will be estimated using a Poisson regression.

Secondary outcome distribution

In addition to testing the suitability of a linear regression for the primary outcome, we also ran analyses to check this assumption for the three secondary outcomes. For brevity, we present the histograms of the outcome distributions for follow-up only.

SDQ: Total Difficulties Score

Figure 4 Follow-up total difficulties score distribution

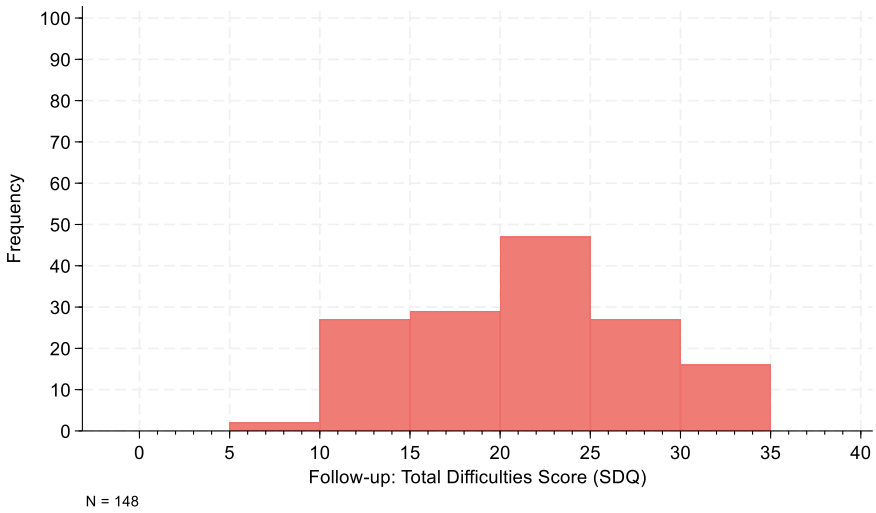
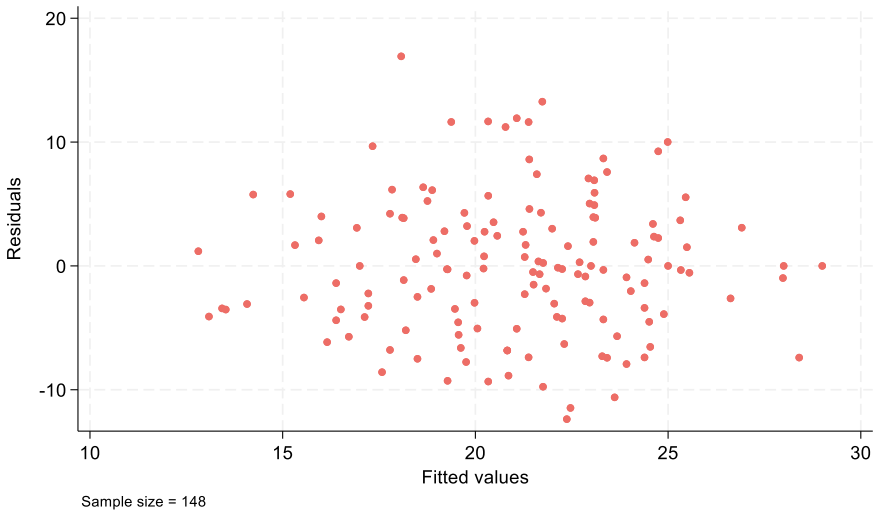


Figure 5 Residual versus fitted plot - Total difficulties score



These graphs indicate that a linear regression is suitable for the analysis of the total difficulties score.

SDQ: Prosocial Behaviour Score

Figure 6 Follow-up Prosocial Behaviour Score distribution

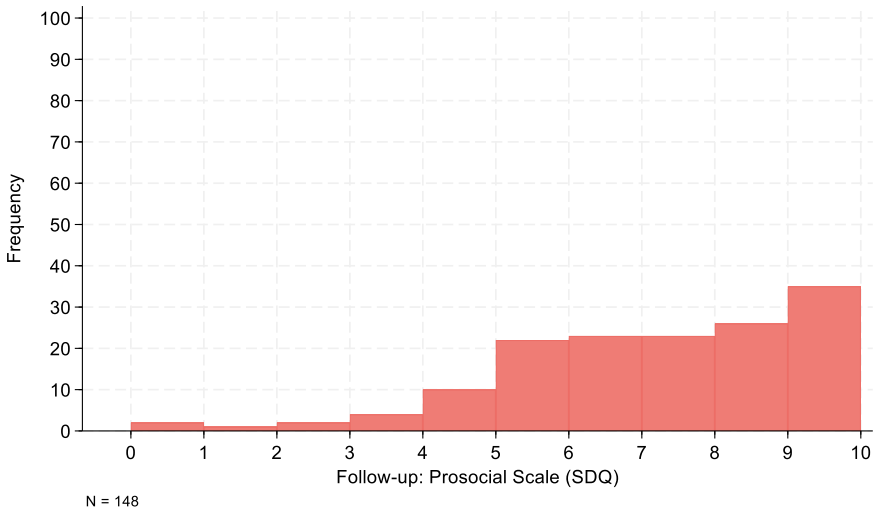
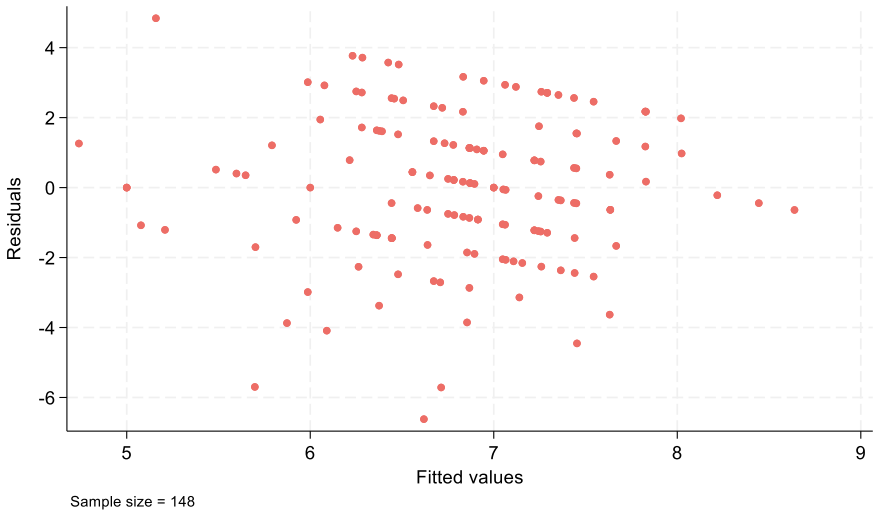


Figure 7 Fitted-versus-residuals plot - Prosocial behaviour score



These graphs indicate that linear regression is suitable for the prosocial behaviour score.

SRSS Total Score

Figure 8 Follow-up SSRS distribution

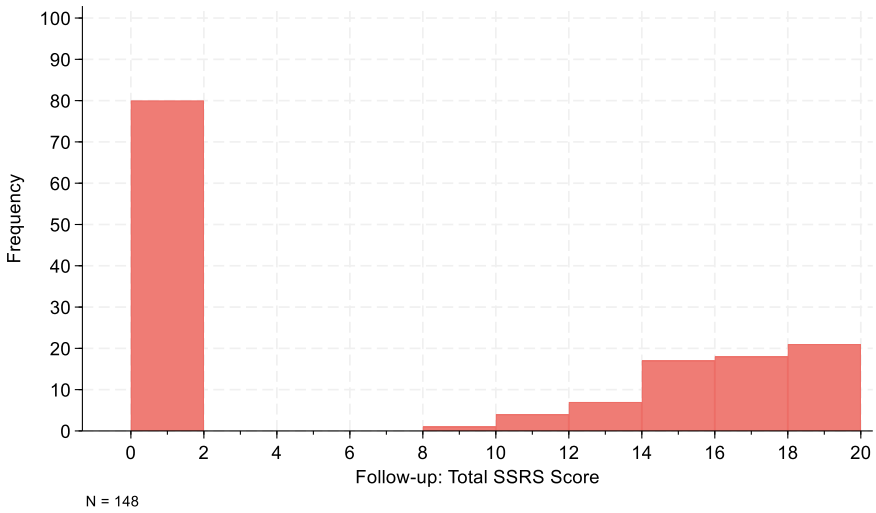
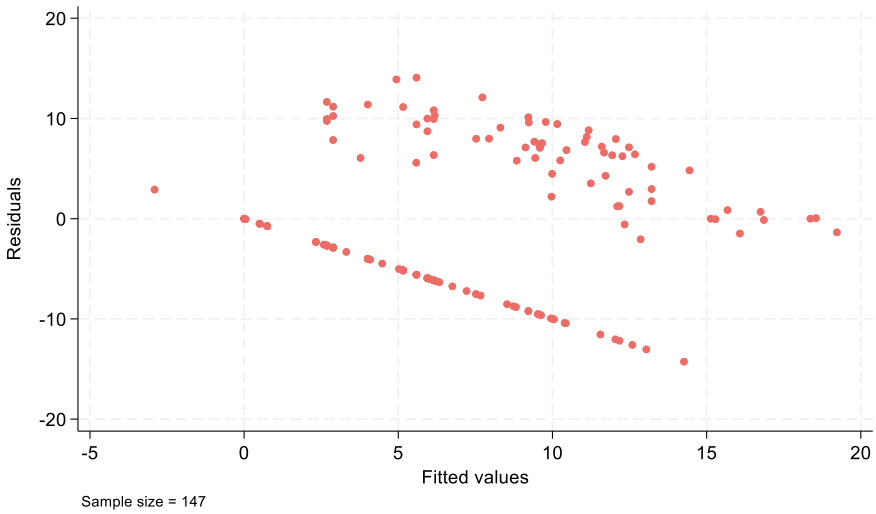


Figure 9 Residual-versus-fitted plot - SSRS



At follow-up, 54% of trial participants had a zero value. These graphs indicate that the analysis of the SSRS should use a Poisson regression model.

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413