Evaluation of the STEER programme

Statistical analysis plan

Evaluating institution: Cordis Bright

Principal investigator(s): Dr Stephen Boxford (PI). Co-PIs: Professor Darrick Jolliffe, Kam Kaur, Suzie Clements, Dr Jade Farrell

YEF statistical analysis plan

Project title[1]	Randomised Controlled Trial of STEER
Developer (Institution)	Salford Foundation
Evaluator (Institution)	Cordis Bright
Principal investigator(s)	Dr Stephen Boxford (PI). Co-PIs: Professor Darrick Jolliffe, Kam Kaur, Suzie Clements, Dr Jade Farrell
SAP author(s)	Professor Darrick Jolliffe, Dr Stephen Boxford, Suzie Clements, Dr Jade Farrell
Trial design	Two-arm, parallel randomised controlled trial with random allocation at the young person level
Trial type	Efficacy study
Evaluation setting	Salford Foundation setting, school, home, community settings
Target group	Young people aged 10-17 and who are at risk of involvement in violent crime because they have an association with peers or family member(s) involved in serious violence, organised

[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).

	crime or gangs and who consent to participate in the programme.
Number of participants	654 young people.
Primary outcome and data source	Reduction in self-reported offending behaviour measured by the Self-reported Delinquency Scale (SRDS) Variety Score (See, Smith & McVie, 2003)
Secondary outcome and data source	Positive relationship between young person and mentor (treatment group) or significant adult (control group) measured by the Social Support and Rejection Scale (SSRS) (Roffman et al. 2000) Improved pro-social values and behaviours measured by the pro-social values subscale in the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 2005) Improved emotional problems measured by the emotional symptoms subscale in the SDQ (Goodman, 2005) Improved behaviours measured by the conduct problems sub-scale in the SDQ (Goodman, 2005) Positive relationships/role models measured by the peer relationships subscale in the SDQ (Goodman, 2005)

SAP version history

Version	Date	Changes made and reason for revision
1.2 [*latest*]	June 2025	Update made to power calculation and MDES based on new information about pre- post- test correlation in relation to SRDS
1.1	6 December 2024	Update made on completion of baseline data collection to show final baseline sample.

1.0 **[*original*]**		

Table of contents

Introduction

This is an efficacy study statistical analysis plan for a two-armed parallel randomised controlled trial (RCT) evaluation of Salford Foundation's STEER programme. The efficacy study included an internal pilot trial started in January 2022 which concluded in May 2023. The trial moved to a full efficacy study in August 2023 and is due to complete in May 2025.

Salford Foundation's STEER programme (STEER) is a six-month intensive mentoring, coaching, family support and case management programme. It pairs young people who are at risk of serious youth violence and child criminal exploitation with a youth worker (mentor). Participants take part in STEER on a voluntary basis. The mentor delivers weekly face-to-face sessions which follow a toolkit of mandatory and optional themed interventions. In addition to these sessions, STEER provides weekly wrap-around case work and support for young people and offers their parents/carers a total of 14 hours of family support to facilitate greater family cohesion.

The key research question of the efficacy study is:

> *"Does a co-designed mentoring, coaching, family support, and case management programme delivered to children and young people with known family members or peers involved in offending behaviour, reduce the likelihood of participant involvement in serious youth violence and future offending or reoffending in comparison to receiving business as usual?"*

The key primary outcome measure for the evaluation will be a reduction in prevalence of self-reported offending behaviours measured by the Self-Reported Delinquency Scale variety scale.[2]

Secondary outcomes include:

- A positive relationship between the young person and a significant adult (e.g., the mentor)
- Improved pro-social values and behaviours (measured by the Strengths and Difficulties Questionnaire (SDQ) pro-social values subscale).

[2] For more information see here: https://res.cloudinary.com/yef/images/v1623145465/cdn/19.-YEF-SRDS-guidance/19.-YEF-SRDS-guidance.pdf

- Improved emotional symptoms (measured by the SDQ emotional symptoms subscale).
- Improved behaviours (measured by the SDQ conduct problems subscale).
- Positive relationships/role models (measured by the SDQ peer relationships problems subscale).

Data for all measures will be collected directly from participants using an online survey administered at baseline and approximately six-months post-randomisation.

Additional research questions include:

1. **Delivery**: Can the STEER programme work under ideal circumstances?
2. **Impact**: a) What is the impact of STEER? b) For whom does STEER work and under what conditions?
3. **Unintended consequences**: a) Does STEER have any unintentional consequences? If so, what are these? b) Do different groups of young people experience these differently?
4. **Iatrogenic effects**: Are there any serious negative effects attributed to STEER on any intended or unintended outcomes?
5. **Mechanisms**: a) How does STEER work to reduce children and young people's involvement in serious youth violence? b) Which factors contribute most to the observed outcomes?

Design overview

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

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

Figure 1 Summary of Efficacy Study design

Trial design, including number of arms	Two-arm parallel randomised controlled trial with random allocation at the young person level
Unit of randomisation	Individual participant
Stratification variables	Not applicable

(if applicable)		
Primary outcome	variable	Reduction in prevalence and variety of self-reported offending behaviours
	measure (instrument, scale, source)	SRDS Variety Score
Secondary outcome(s)	variable(s)	A positive relationship between the young person and a significant adult (e.g., the mentor) Improved pro-social values and behaviours Improved emotional symptoms Improved behaviours Positive relationships/role models
	measure(s) (instrument, scale, source)	A positive relationship between the young person and a significant adult (e.g., the mentor) measured by the Social Support and Rejection Scale (SSRS) (Roffman et al. 2000) Improved pro-social values and behaviours measured by the pro-social behaviour sub-scale in the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 2005) Improved emotional symptoms measured by the SDQ emotional symptoms sub-scale (Goodman, 2005) Improved behaviours measured by the SDQ conduct problems sub-scale (Goodman, 2005) Positive relationships/role models measured by the Peer relationships problem sub-scale in the SDQ (Goodman, 2005)
	variable	Reduction in prevalence and variety of self-reported offending behaviours

Baseline for primary outcome	measure (instrument, scale, source)	SRDS Variety Score
Baseline for secondary outcome	**variable**	Improved pro-social values and behaviours Improved emotional symptoms Improved behaviours Positive relationships/role models
	measure (instrument, scale, source)	Improved pro-social values and behaviours measured by the Strengths and Difficulties Questionnaire (SDQ) pro-social values sub-scale (Goodman, 2005) Improved emotional symptoms measured by the SDQ emotional symptoms sub-scale (Goodman, 2005) Improved behaviours measured by the SDQ conduct problems sub-scale (Goodman, 2005) Positive relationships/role models measured by the measured by the SDQ peer relationships problem subscale (Goodman, 2005)

Sample size calculations

Originally, we determined the final required participant sample size a priori, conducting Power Calculations in line with YEF guidance which suggested a total sample of 654 young people (327 per group) over the pilot trial and efficacy study would allow a statistically significant result to be identified (Power=0.80, two tailed, P<.05) for a reduction of involvement in offending of 11%. We recognised that to account for attrition STEER would need to recruit a greater number of young people to reach a final sample size of 654 by the end of the evaluation.

Our original approach was conservative and in line with Lipsey and Wilson (2001) who state that ½ d=r, which is in turn equivalent to the difference in proportions. Figure 2 shows that if we suggest that 30% of the young people that STEER does not work with commit violence compared to 20.5% of the young people that STEER does work with committing violence (equivalent to a Cohen's d=.19) a total sample of 654 (327 in each group) would be needed to detect a statistically significant result (Power=.80), in a two-tailed test (p<.05). This level of Cohen's d was selected because it is conservative and is about equivalent to a 10-11%

difference which is in line with a weighted average effect size of mentoring programmes, based on comparisons of 18 studies in a meta-analysis of mentoring and offending using a random effects model (d=.21, 95% confidence interval .07 to .34) presented by Jolliffe and Farrington, 2008.[3]

Figure 2 presents the results of our power analysis at the protocol stage. It shows that in line with our conservative approach; we used a pre-test/post-test correlation of 0. This is because at the time we had no reason to believe based on data collected during the pilot trial that the variance would be different between the treatment and control group. However, we knew that inclusion of a pre-test as a covariate in impact analyses helps to explain (error) variance in the post-test and improves the likelihood of uncovering programme impacts by reducing the standard error of the impact estimate. It was difficult to estimate what the pre-test/post-test correlation would be as this depended on unknown sample characteristics and the characteristics of the measure under investigation (the SRDS when used in a sample similar to STEER, i.e., those who are known to have peers or family members involved in offending behaviour). The greater the estimated pre-test/post-test correlation the lower the MDES and the smaller the sample needed to detect this. In practice however, if the pre-test/post-test correlation changes from 0.0 to 0.4 the MDES for a sample size of 500 decreases from .25 to .23.

In line with good practice, we have conducted power analysis based on our final sample once all randomisation has been completed. This power analysis is based on the numbers involved in the trial at baseline. This is also presented in Figure 2. As part of this calculation, we have included a pre- post-test correlation of 0.5. 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 variety score as a continuous variable in analysis as opposed to a dichotomous variable.

Our power analysis shows an MDES of 0.186 based on baseline data

SPSS 25 was used for calculations for the original protocol. PowerUp! (Dong, N. and Maynard, R. A., 2013) has been used for the calculations once all randomisation and baseline data had been completed.

[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.

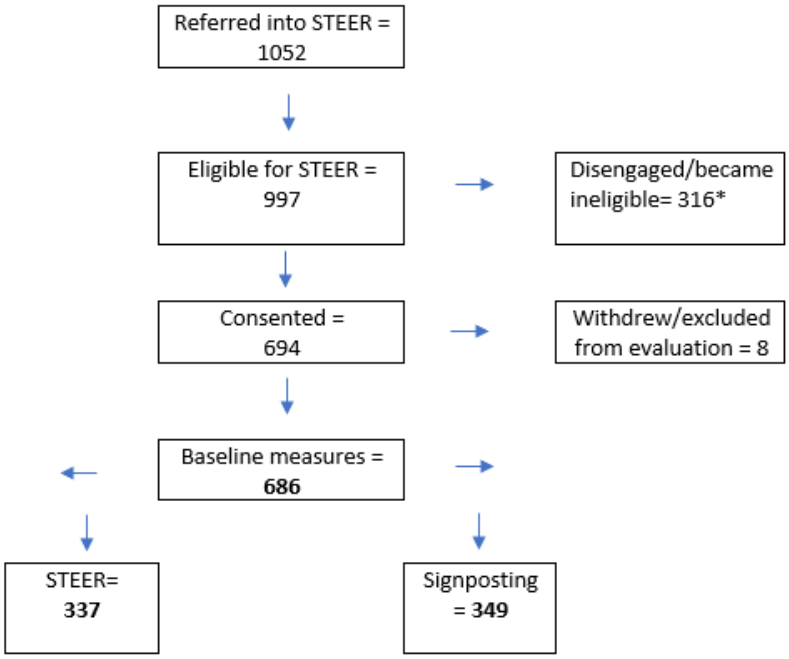
Figure 2: Sample size calculations

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		0.19	0.186
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 by end of the trial	Intervention	327	337
	Control	327	349
	Total	654	686

Recruitment and baseline data collection was completed in October 2024. The cumulative T1 sample at the end of this period was 686.

337 young people were allocated to the intervention group, and 349 to the control group. This is broadly in line with the intended 1:1 allocation ratio. More information can be seen in Figure 3 below.

Figure 3: Consort flow diagram for baseline data collection



* These are young people who disengaged or became ineligible because they: refused the service, had a parent who refused the service, moved out of area, were uncontactable, had too many services involved, or were deemed too complex due to an escalation in risk following the acceptance of referral.

Analysis

Overview

The following presents the analytical approach for the Efficacy Study. We currently do not know the structure of the complete data-set (for example, extent of or categories of missing data etc.) that we will receive as part of the Efficacy Study. This should be considered when interpreting the following analytical plan.

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

Based on our understanding of the literature our analysis will be conducted using General Linear Models controlling for baseline measurement of the outcome variable. An intention to treat approach will be used with all models unless otherwise stated. This in line with YEF Guidance (YEF, 2021).

The primary outcome for the evaluation of STEER is a reduction in prevalence of self-reported offending behaviours between baseline and six months as measured by the SRDS variety score.

The secondary outcomes that we are investigating are that young people receiving STEER compared to the business as usual (control) group have (See Figure 1 for more information about measures):

- A positive relationship with their mentor.
- Improved pro-social values and behaviours.
- Improved emotional symptoms.
- Improved behaviours
- Positive relationships/role models.

The General Linear Model will include the appropriate baseline outcome measure and the treatment dummy variable. For example, for the analysis of the primary outcome measure (SRDS Variety Score) the baseline SRDS variety score measure will be included and whether the individual received STEER or not

The purpose of these analyses is to estimate the difference between the STEER arm and the business-as-usual arm for each of the primary and secondary outcomes.

Variable transformation

It is possible that the baseline and outcome variables may be skewed. Skew will be assessed using the traditional criteria based on their distribution (i.e., skews of greater or equal to 1.0 or less than or equal to -1.0). Arguably, it is more desirable to use a generalised liner model (GLM) for the appropriate modelling of non-normally distributed variables (e.g., Akram et al., 2023), than it is to transform the data.

The analytic approach has been developed a priori and will be conducted with SPSS.

The syntax for all analysis will be provided once it has been developed after all data has been collected.

Primary outcome analysis

The primary outcome is a reduction in offending measured by the SRDS Variety Score. All young people will have completed the SRDS questionnaire before randomisation and again at around six months post-randomisation.

The SRDS Variety Score is measured on a scale from 0-19 with 0 indicating the young person has not reported any of the 19 forms of delinquency/offending behaviour and a score of 19 indicating that they have undertaken all forms of delinquency/offending behaviour.

We will be using a General Linear Model, repeated measures design (assuming normality) or a generalized linear model, repeated measures design if the SRDS Variety score is non-normally distributed. We will include a treatment by outcome interaction term in the analysis.

This analysis is designed to evaluate the differences in SRDS Variety Score between those in STEER and those who received business as usual (i.e. the control group).

This analysis will be conducted in SPSS.

Secondary outcomes analysis

Our approach to analysis of secondary outcomes will mirror the approach outlined above for the analysis of the primary outcomes. The secondary outcomes are:

- A positive relationship between the young person and mentor measured by the Social Support and Rejection Scale (SSRS).
- Improved pro-social values and behaviours measured by the SDQ pro-social values subscale.
- Improved emotional symptoms measured by the SDQ emotional symptoms subscale.
- Improved behaviour measured by the SDQ conduct problems subscale.
- An increase in positive relationships/role models measured by the measured by the SDQ peer relationships problem subscale.

All SDQ subscales contain 5 items and are measured on scales from 0 to 10. For the prosocial values subscale high scores are desirable (e.g., greater prosocial values), but for the other subscales (e.g., emotional symptoms subscale, conduct problems subscale, peer relationship subscale) high scores are not desirable (e.g., greater emotional problems, greater conduct problems, poorer peer relationships).

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

We will be using general linear models, repeated measures design (assuming normality) or generalized linear models, repeated measures design if the particular subscale is non-normally distributed. This will be determined once all data has been collected. We will include a treatment by outcome interaction term in the analysis.

This analysis is designed to evaluate the differences in prosocial values, emotional symptoms, behaviour, peer relationships and positive relationships between the young person and mentor between those in STEER and those who receive business-as-usual.

This analysis will be conducted in SPSS.

Subgroup analyses

The subgroup analyses we will consider undertaking will be exploratory in nature. Before considering undertaking sub-group analyses we will assess whether these would be sufficiently powered based on the data we have collected. We will assess whether we are likely to have a sufficient number of young people in the groups by undertaking a power analysis after all data has been collected before proceeding with these analyses. We will explore the following analyses:

- **Race equity, equality, diversity and inclusion.** We will consider whether STEER was equally effective for those from minoritized backgrounds compared to those from White backgrounds. Given the limited knowledge about the effectiveness of interventions with those from minoritised backgrounds we would propose to conduct an exploratory analysis to consider whether the intervention was equally effective for those of Black (e.g., Black Caribbean/Black African) backgrounds to those of White backgrounds. This would likely be an underpowered analysis so caution would be taken interpreting the results.
- **Reduced offending as measured by police data.** If we are able to access the right kind of police data we will explore whether STEER had an impact on reducing offending over and above that reported in the business-as-usual group.

These analyses would be conducted using a general linear model, repeated measures design (assuming normality) or a generalized linear model. We will include a treatment by outcome interaction term in the analysis. As stated, we will only proceed with subgroup analyses where Power Calculations suggests the analyses will be sufficiently powered.

Further analyses

We will evaluate the extent to which positive relationships between the young person and mentor (treatment group) or significant adult (control group) influenced the primary outcome over and above the impact of STEER through the SSRS.

We are proposing conducting this analysis because the theory of change suggests that a mechanism for STEER is that it has its effect through an increase in a positive relationship with a mentor.

Analysis will be conducted using a General Linear Model, repeated measures design (assuming normality) or a Generalized Linear Model.

Interim analyses and stopping rules

As part of the pilot trial, we analysed the completeness, reliability and validity of outcomes questionnaires (including the measures of outcomes described above). We did this by conducting regular data quality audits including exploring: percentages of scale item completeness, scale means, standard deviations and skew as well as conducting Cronbach Alpha testing for scale reliability and correlation analysis to test theoretical validity. We will continue this approach for the duration of the Efficacy Study. Our interim analyses as part of the pilot trial did not, and will not, include a comparison between control group and treatment group data nor analyses of impact.

The trial will stop if Salford Foundation, YEF and Cordis Bright decide that STEER is unable to recruit a sufficient number of participants. Recruitment rates will be monitored against modelled targets regularly (monthly) and reviewed bi-monthly (at a minimum) as part of project group meetings. Any decision about stopping will be made in discussion with YEF and Salford Foundation colleagues.

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

Longitudinal follow-up analyses

Other than assessing change in outcomes between baseline and follow-up as described above, there will be no other longitudinal follow-up analyses as part of the STEER evaluation.

Imbalance at baseline

We will produce a table of baseline descriptive characteristics for all young people before they were randomised and for those analysed. The baseline characteristics will include age, sex, ethnicity, and the relevant outcomes (SRDS Variety Score and SDQ subscales). We will report counts (including the numerator and denominator) and percentages in each category. Any differences will be discussed in the final evaluation report.

Missing data

We will assess missing data before analysis. We will follow YEF guidance as appropriate and report on both: (1) the number of complete cases, and (2) the extent and pattern of missingness in the data. We will also attempt to establish the missing mechanism (i.e. what variables in the data are predictive of non-response). We will explore this through logistic regression models where the presence of missing data will be modelled with additional information that may be predictive of missingness. We will conduct this analysis in line with YEF guidance and will also discuss the types of missing data in the final report as per Table 1 and using the flow chart in Figure 2 in the YEF analysis guidance. For more information see:

https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf Extent of and reasons for missing data will be assessed and summarised in the final report.

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

Compliance

Overall compliance for the purposes of the efficacy study will be met when young people have been randomised and allocated into the treatment or control group. This is in line with the intention to treat approach specified in the YEF Statistical Analysis Guidance (YEF, 2021).

We will explore model compliance if Power Calculations suggests the analysis will be sufficiently powered. This will explore what level of dosage was associated with a desirable outcome on the SRDS. For example, does attending 75% of STEER mandatory sessions result in a similar impact as attending all sessions? This analysis will be conducted using a general linear model, repeated measures design (assuming normality) or a generalized linear model. We will include a treatment by outcome interaction term in the analysis.

A note on intra-cluster correlation

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

Presentation of outcomes

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

Hedges' $g = (x1 - x2) / \sqrt{((n1-1)*s12 + (n2-1)*s22) / (n1+n2-2)}$

where:

x1, x2: The sample 1 mean and sample 2 mean, respectively

n1, n2: The sample 1 size and sample 2 size, respectively

s12, s22: The sample 1 variance and sample 2 variance, respectively

With a sample of greater than 20 there is limited difference with Cohen's d. However, if the standard deviations between the treatment and comparison group are different, we would

propose to use Glass' delta, which only uses the control group's standard deviation (Lipsey & Wilson, 2001).

The confidence interval for the Hedge's g statistic is:

$g \pm \phi - 1\ (1-(\alpha/2))gse$

where:

$\phi - 1$ = the percent point function of the normal distribution

gse= the standard error of the g statistic

$= \sqrt{(n1+n2)/n1n2 + g^2/2(n1+n2)}$

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

Goodman, R., 2005. *The Strengths and Difficulties Questionnaire.* Available at: https://www.sdqinfo.org/a0.html [Accessed 17/8/23].

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.

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. (2021) *Practical meta-analysis.* Thousand Oaks, CA: Sage.

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.

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

Youth Endowment Fund, 2021. *Analysis Guidance.* Available at: https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf [Accessed 20 /10/23].

Youth Endowment Fund, 2021. *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].