STATISTICAL ANALYSIS PLAN

The Reach Programme

Sheffield Hallam University

Principal investigators: Anna Stevens, Charlotte Coleman, Ben Willis

YOUTH ENDOWMENT FUND

Statistical analysis plan: The Reach Programme

Evaluating institution: Sheffield Hallam University
Principal investigator(s): Anna Stevens, Charlotte Coleman and Ben Willis

Project title[1]	The Reach Programme
Developer (Institution)	VRU for Leicester, Leicestershire and Rutland, Leicester City Council, Leicestershire County Council
Evaluator (Institution)	Sheffield Hallam University
Principal investigator(s)	Anna Stevens, Charlotte Coleman, Ben Willis
SAP author(s)	Anna Stevens, Sean Demack
Trial design	Blocked Randomised Controlled Trial (RCT) design. Within each block (i.e. school), referred young people (YP) will be randomised to receive the Reach programme or BAU control (i.e. two arms).
Trial type	Efficacy with internal pilot
Evaluation setting	School
Target group	Children and YP aged 11 to 16 years old who are at risk of suspension or who are persistently absent from school, and where there are concerns about future involvement in anti-social behaviour and crime as both a victim or perpetrator.
Number of participants	12 schools, 600 YP

Primary outcome and data source	Externalising score (Teacher report SDQ)
Secondary outcome and data source	• Internalising score (Teacher report SDQ) • Externalising score (Self report SDQ) • Internalising score (Self report SDQ) • Prosocial behaviour score (Teacher report SDQ) • Prosocial behaviour score (Self report SDQ) • Variety of delinquency score (SRDS) • Volume of delinquency score (SRDS) • Number of suspensions (source: Leicester City Council and Leicestershire County Council) • School attendance (source: Leicester City Council and Leicestershire County Council) • Offending data (source: VRU Leicester)

SAP version history

Version	Date	Changes made and reason for revision
1.0	09/05/24	NA

Table of contents

Contents

Introduction

The Reach Programme is targeted intervention aimed at YP (young people) aged 11-16 in secondary schools (years 7-11) who are at risk of suspension (i.e. they have carried out behaviour in their school that would normally qualify for a suspension) or who are persistently absent from school, have 3 indicators of vulnerability (e.g. looked after, domestic violence or substance misuse in the home) and where there are concerns about future involvement in anti-social behaviour and crime as both a victim or perpetrator. YP are referred to the Reach Programme directly from schools. A team of 10 Youth Workers working across the 12 schools deliver one to one sessions to YP over a period of 6 months, covering a set of core components as follows:

- Relationship Building
- Understanding behaviour
- Social Skills training
- Confidence, Wellbeing and Resilience
- Positive Family, Peer, and Community Relationships
- Identifying and Achieving Aspirations
- Recreational Activity

The Reach evaluation is designed as an efficacy trial with internal pilot. The internal pilot phase of the evaluation is now complete, and the efficacy phase began in January 2024 across the same 12 schools that were recruited for the pilot phase. Whist the sample of 12 schools remains the same, the YP recruited to the efficacy phase are different to those recruited at the pilot phase. Please see the Reach Protocol for a more detailed description of the intervention and associated documents (referral form, Theory of Change and Logic Model).

Design overview

Table 1: Design overview

Trial design, including number of arms		Two-arm randomised control trial
Unit of randomisation		Individual participant
Stratification variables (if applicable)		School (to balance allocation within schools)
Primary outcome	variable	Behavioural difficulties
	measure (instrument, scale, source)	Teacher report SDQ externalising score [0 to 20]
Secondary outcome(s)	variable(s)	• Emotional regulation and peer relationships (Teacher report SDQ) • Behavioural difficulties (Self report SDQ) • Emotional regulation and peer relationships (Self report SDQ) • Prosocial behaviour (Teacher report SDQ) • Prosocial behaviour (Self report SDQ) • Variety of delinquency • Volume of delinquency • Number of suspensions • Attendance at school (%) • Level of offending
	measure(s) (instrument, scale, source)	• Teacher report SDQ internalising score [0 to 20] • Self report SDQ externalising score [0 to 20] • Self report SDQ internalising score [0 to 20] • Teacher report prosocial score [0 to 10] • Self report prosocial score [0 to 10] • Variety of delinquency (SRDS) • Volume of delinquency (SRDS) • Number of suspensions • Attendance at school (%) • Level of offending

Baseline for primary outcome	**variable**	Behavioural difficulties
	measure (instrument, scale, source)	Teacher report SDQ externalising score [0 to 20]
Baseline for secondary outcome	**variable**	• Emotional regulation and peer relationships (Teacher report SDQ) • Behavioural difficulties (Self report SDQ) • Emotional regulation and peer relationships (Self report SDQ) • Prosocial behaviour (Teacher report SDQ) • Prosocial behaviour (Self report SDQ) • Variety of delinquency • Volume of delinquency • Number of suspensions • Attendance at school (%) • Level of offending
	measure (instrument, scale, source)	• Teacher report SDQ internalising score [0 to 20] • Self report SDQ externalising score [0 to 20] • Self report SDQ internalising score [0 to 20] • Teacher report prosocial score [0 to 10] • Self report prosocial score [0 to 10] • Variety of delinquency (SRDS) • Volume of delinquency (SRDS) • Number of suspensions • Attendance at school (%) • Level of offending

Sample size calculations overview

The power calculations assume an ITT analyses with 100% complete baseline/outcome data and that the only systematic difference between the Reach group and the control group is their group membership (all other difference are random). Table 2 below presents the sample size calculations, showing firstly the estimates presented in the protocol from which the sample size for the study was drawn (600 YP) to achieve an MDES of 0.18 – 0.19. This was determined a priori using correlation estimates (0.50 to 0.60), drawn from the EEF Adventure Learning Trial (Willis et al., 2023), and drawing on the Reach baseline pilot data for the ICC (0.13). The sample size of 600 YP was discussed with the delivery partner and the timescales to achieve this sample size were agreed, also taking into account practical considerations in terms of delivery. Subsequent to this the internal pilot phase has now been completed, giving an updated participant level correlation of 0.32 (calculated from the full set of pilot data), which would achieve an MDES of 0.21 for the same sample size (Table 2).

Table 2: Sample size calculations

Sample = 600 CYP across 24 clusters (25 CYP per cluster)		**Protocol**	**Randomisation: drawing on Reach pilot data for ICC and correlation estimates***
Minimum Detectable Effect Size (MDES)		0.18-0.19 sds	0.21 sds
Pre-test/ post-test correlations	level 1 (participant)	0.50 to 0.60	0.32 (R^2=0.10)
	level 2 (cluster)	n/a[1]	n/a[1]
Intracluster correlations (ICCs)	level 1 (participant)	n/a[2]	n/a[2]
	level 2 (cluster)	0.13	0.13
Alpha[1][1]		0.05	0.05
Power		0.80	0.80
One-sided or two-sided?		Two	Two

*Internal pilot phase was completed subsequent to protocol submission and is drawn on to inform the SAP

The power calculations were undertaken using a formula and checked using the PowerUp software (Dong et al., 2013, sheet BIRA1_1r)[2]. For the Reach evaluation, this is a blocked RCT

[2] Dong, N., & Maynard, R. A. (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), 24-67.doi: 10.1080/19345747.2012.673143. https://www.causalevaluation.org/power-analysis.html

design such that in each 'block' (school), YP are randomly allocated to Reach programme, or to the control group. The power analysis was conducted in the following way:

$$MDES_{RCT(Blocked)} \sim M_{J-L-1} \sqrt{\frac{ICC_2(1-R_2^2)\varphi}{J} + \frac{(1-ICC_2)(1-R_1^2)}{P(1-P)Jn}}$$

Where:

- J=number of blocks (schools)
- n=number of individuals per block (YP per school)
- L= number of (block level) covariates to be used in the impact analyses (this might include a variable used to stratify or minimise in randomisation)
- P = the proportion of individuals assigned to the intervention in each location/block - which would be 0.5 within a balanced design.
- M_{J-L-1} is the group effect multiplier … value from the t-distribution for 2-tailed test with alpha=0.05 & beta=0.80, equal variances assumed with J-L-1 degrees of freedom.
- ICC_2 is the cluster (location/block) level ICC; the proportion of variance in the outcome between blocks.
- R_2^2 = proportion of between cluster/block level variance that is reduced by covariate(s) - block level explanatory power.
- R_1^2 = proportion of within-block variance that is reduced by covariate(s) - participant level explanatory power.
- φ = Treatment effect heterogeneity (which is usually set to be zero for efficacy trials – larger effectiveness trials might be powered to detect this).

If we assume zero treatment effect heterogeneity and a balanced design … P=0.5 and φ = 0, equation 1.1 simplifies to equation 1.2:

$$MDES_{RCT(Blocked)} \sim 2M_{J-L-1} \sqrt{\frac{(1-ICC_2)(1-R_1^2)}{Jn}}$$

This removes R_2^2 from the equation (because φ = 0) and so only explanatory power at the YP level influences the statistical sensitivity of the blocked-RCT design.

The impact of the ICC on sensitivity is the opposite of what is seen with clustered RCT design that randomise at the cluster level. For clustered designs with randomisation at the cluster level (e.g. school), higher ICCs are associated with higher MDES estimates. For blocked-RCT

designs with randomisation at the individual level in each 'block' (e.g. school), higher ICCs are associated with lower MDES estimates. This is because the blocked-RCT design fixes things so that the cluster (or block) level is completely accounted for – i.e. the Reach intervention and BAU control groups are randomised in exactly the same locations (or blocks). Therefore, as the cluster-level variance in the outcome increases proportionally (i.e. as ICCs increase), an increasing quantity of random 'noise' is completely accounted for and this results in increased statistical sensitivity (as seen with decreasing MDES estimates).

As ICC_2 increases, (1-ICC_2) and (therefore) the MDES estimate decreases. So, to estimate the smallest effect size that a blocked design could detect as statistically significant (P,0.05, two-tailed) with a statistical power of 0.80 or higher [MDES], the following details are needed:

- J=number of schools = 12 at pilot; 12 at stage 2 ~ 24 in all. n= CYP per school = 25. L= number of (block level) covariates =0
- M_{J-L-1} is the group effect multiplier … value from the t-distribution for 2-tailed test with alpha=0.05 & beta=0.80, equal variances assumed with J-L-1 (24-0-1=23) degrees of freedom.
- ICC_2 is the cluster (location/block) level ICC; the proportion of variance in the outcome between blocks: Unknown, estimated as 0.13 from baseline pilot data.
- R_1^2 = proportion of within-block variance that is reduced by covariate(s) - participant level explanatory power. This was estimated at between 0.50 and 0.60 based on findings from Adventure Learning (Willis et al. 2023). Also presented in table 2 are calculations based on a participant level correlation of 0.32 calculated from the full set of pilot data for this trial that is now available.

Analysis

The methods of analysis described here were chosen prior to data collection. Multi-level linear regression models will be constructed that acknowledge that pupils are clustered in schools. Specifically, schools will be included as random effects within two-level random intercepts multilevel models. Elff, M et al. (2021) justify the use of this approach with a smaller number of clusters. The Reach trial is an efficacy trial and therefore conditional inference only will be made, we will not be generalising beyond the sample of schools included in the study. All multi-level models will be conducted in STATA version 17. In each of these two models, the follow up teacher report SDQ externalising score will be the outcome variable and the trial arm (1=Reach or 0=Control) included as the key pupil-level explanatory variable. Baseline teacher report SDQ externalising score will be includes as covariates at both pupil and school levels. An intention to treat (ITT) approach will be taken that includes all YP randomly allocated to the Reach or to the control group regardless of whether the Reach programme was engaged with. The ITT approach best preserves randomisation (and therefore the

strength of internal validity) and will provide the most robust estimate of the causal impact of the Reach programme. The headline ITT analysis will combine data from the internal pilot and efficacy stages but follow on sensitivity analyses will explore impact at each stage. The impact of Reach will be estimated by converting the model coefficient for the trial arm variable into Hedges' g effect sizes using the equation below, where T is the treatment mean, C is the control mean, δ^2_{sch} is the school level variance and δ^2_{pup} is the pupil level variance:

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

For the primary outcome analysis and follow-on exploratory analyses, statistical uncertainty will be expressed as standard errors of multilevel model coefficients and use of 95% confidence intervals.

Primary outcome analysis

The primary outcome measure for the Reach trial is behavioural difficulties as measured by the externalising score contained within the teacher version of the SDQ (Goodman, 2001). The SDQ, teacher version is a 25-item scale used to assess behaviour in the school context in 4–16-year-olds. This consists of 5 subscales; conduct problems, hyperactivity scale, emotional problems scale, peer problems scale and prosocial scale. The externalising score measures "behavioural difficulties" and is the sum of the conduct and hyperactivity scales. This links well with the Theory of Change in terms of improved behaviour management, reduction in negative behaviours in school and reduction in problem behaviours both in and out of school. This is also one of YEF's core outcome measures. The primary research question for the Reach trial is set out below:

RQ1 What is the difference in behavioural difficulties measured using the externalising score from the teacher report Strengths and Difficulties Questionnaire (SDQ) between the intervention group, when compared to a 'business as usual' control?

Table 3 below provides an example analysis model for RQ1.

Table 3: Example analysis model RQ1

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*
Empty model			Follow up teacher report SDQ externalising score
ITT sample (RQ1)	Group (1=Reach participants school, 0=Control group participants) Baseline teacher report SDQ externalising score (centred around the mean)	Mean school-level baseline teacher report SDQ externalising score (centred around school level Grand mean)	

Secondary outcome analysis

The variables to be employed for the secondary outcome analysis derive from the teacher report SDQ, the self-report SDQ, the self report delinquency scale (SRDS) and administrative data on suspensions, attendance and offending data as described below and presented in table 1. The teacher report SDQ internalising score is the sum of the emotional and peer problems scales and will form one of the secondary outcomes which links to improved emotional regulation and increased self-esteem and emotional well-being, and also increased network of positive peers. The teacher report prosocial score forms a further secondary outcome and links to improved social skills. The self-report SDQ (completed by the YP) externalising score, internalising score and prosocial score will be used as secondary outcomes, linking to the LM in the same way as the teacher version, and providing validation of findings from the primary outcome.

The self-report delinquency scale (SRDS) will also be used as a secondary outcome. This measures delinquent behaviour by assessing the frequency and severity of any delinquent acts committed. This fits with the LM in terms of reduction of negative behaviours at school and reduction in suspensions or problem behaviours, and improved attendance at school.

A further three outcome measures will be analysed from administrative data; suspensions, attendance and offending behaviours. Using this administrative data provides further validation to the self report measures in terms of attendance and offending behaviours which link to the LM as described above. Using data on suspensions as a secondary outcome links directly to the LM in terms of reduction in suspensions.

Attendance data is reported as a percentage of attendance at school by the YP. This was collected at the point of referral from the school, and collected again from the schools at the

same time as the teacher report SDQ outcome measure. This data will be verified with the councils, who will then share the data with SHU. Data on suspensions (number of previous suspensions) was also collected from schools at the point of referral. This data is being collected directly from the local councils by the delivery partner at the point of closure to the Reach Programme, and the equivalent time period for the control group and shared with SHU to be matched into the final dataset using YP name/school/date of birth to achieve accurate matching. The secondary outcome analysis research questions are set out below:

- **RQ2** What is the difference in emotional regulation and peer relationships measured using the internalising score of the teacher report SDQ between the intervention group, when compared to a 'business as usual' control?
- **RQ3** What is the difference in behavioural difficulties measured using the externalising score from the self report Strengths and Difficulties Questionnaire (SDQ) between the intervention group, when compared to a 'business as usual' control?
- **RQ4** What is the difference in emotional regulation and peer relationships measured using the internalising score of the self report SDQ between the intervention group, when compared to a 'business as usual' control?
- **RQ5** What is the difference in prosocial behaviour measured using the prosocial score of the teacher report SDQ between the intervention group, when compared to a 'business as usual' control?
- **RQ6** What is the difference in prosocial behaviour measured using the prosocial score of the self report SDQ between the intervention group, when compared to a 'business as usual' control?
- **RQ7** What is the difference in offending behaviours measured using "variety of delinquency score" from the self-report delinquency scale (SRDS) between the intervention group, when compared to a 'business as usual' control?
- **RQ8** What is the difference in offending behaviours measured using "volume of delinquency score" from the self-report delinquency scale (SRDS) between the intervention group, when compared to a 'business as usual' control?
- **RQ9** What is the difference in attendance at school measured using administrative data between the intervention group, when compared to a 'business as usual' control?
- **RQ10** What is the difference in number of suspensions measured using administrative data between the intervention group, when compared to a 'business as usual' control?
- **RQ11** What is the difference in offending behaviours measured using administrative data between the intervention group, when compared to a 'business as usual' control?

Analysis of secondary outcomes will be conducted in the same way as for the primary outcome, table 4 below provides an example analysis model for RQ7.

Table 4: Example analysis model RQ7

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*
Empty model			Follow up variety of delinquency score (SRDS)
ITT sample (RQ7)	Baseline variety of delinquency score (SRDS) (centred around the mean) Group (1=Reach participants school, 0=Control group participants)	Mean school-level variety of delinquency score (SRDS) (centred around school level Grand mean)	

Subgroup analyses

The subgroup analysis was specified a priori and reflects the description of this analysis in the protocol. Exploratory analyses of sub-group will examine evidence of differential impact relating to YP ethnicity, gender and age of the YP. Sub-group analysis by ethnicity will further inform the RQs set out in the IPE which explore how far the intervention is reaching YP from ethnic minority backgrounds, and will allow an exploration as to whether outcomes differ by ethnicity. In terms of gender of the YP, the qualitative findings from the feasibility study showed that young males recruited to the Reach Programme responded well to a predominately female YW team, suggesting that the gender of the YW was less important in terms of responsiveness of male participants. Exploring this factor in the subgroup analysis will further inform this research question by allowing an exploration of any differences in outcomes between males and female participants. Findings from the feasibility study suggested that schools were keen to refer YP at the earliest point possible in their lives given that the Reach Programme is intended as a preventative intervention. Analysis of outcomes by age will allow insight into whether the Reach Programme shows any difference in outcomes for younger participants. The research questions for the sub-group analysis are set out below:

- **RQ12** Are any differences in the primary outcome (externalising score teacher report SDQ) observed with regards to the ethnicity of the YP?
- **RQ13** Are any differences in the primary outcome (externalising score teacher report SDQ) observed with regards to the sex of the YP?
- **RQ14** Are any differences in the primary outcome (externalising score teacher report SDQ) observed with regards to the age of the YP?

The final categories for ethnicity and age will be decided upon examination of the final achieved sample to inform the categories that contain sufficient numbers for analysis. Detailed data on these subgroups is being collected as part of the monitoring data, and will be matched into the final dataset.

As illustrated in Table 5 (using RQ13 as an example), subsample analyses will first be undertaken by including two additional variables to the model; a main effect term (i.e. ethnicity, gender or age) and an interaction between this and group membership. These analyses will directly examine evidence of differential impact for Reach. If the interaction term is found to be statistically significant, follow-on analyses will undertake separate impact analyses for the subsamples.

Table 5: Example analysis model RQ13

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*
Empty model			Follow up variety of delinquency score (SRDS)
ITT sample (RQ7)	Group (1=Reach participants school, 0=Control group participants) Baseline variety of delinquency score (SRDS) (centred around the mean) Sex (1=Female, 0=Male) Sex*Group interaction	Mean school-level variety of delinquency score (SRDS) (centred around school level Grand mean)	

Imbalance at baseline

Our examination of imbalance at baseline will focus on the baseline measures collected (described in table 1) and YP characteristics including age, gender and ethnicity. A table of baseline characteristics will be presented, showing counts and percentages for categorical variables, and means and standard deviations for continuous variables. These analyses will provide an indication of imbalance at baseline following randomisation. Effect sizes will be calculated from the descriptive statistics generated from scale variables and then used to determine where sensitivity analysis is needed. Austin (2009) suggests that a standardised

difference of 0.1 denotes meaningful imbalance, this is especially relevant if the variables in which imbalance is seen are highly predictive of outcomes (Ho et al., 2007). If imbalance at baseline is observed, a sensitivity analysis will be included to address this by including the variable (with observed imbalance) as a covariate as an efficient method for achieving better balance (Hewitt & Togerson, 2006).

Missing data

The baseline and ITT samples will be compared to help illustrate the impact of missing data for the primary outcome variable only. A summary of known reasons for missing data will be presented (e.g. YP moved school/YP excluded from school). This will firstly be done descriptively by tabulating missing cases across the categories of variables included in the ITT analysis. Reasons for any missingness will be summarised and we will examine whether missingness is associated with school and/or pupil-level covariates for example; baseline data. A multi-level logistic regression model (1=in ITT model; 0=not in ITT model) will examine whether missingness is associated with school or pupil-level covariates (including age/gender/ethnicity of the YP).

If over 5% of cases in the baseline sample are missing from the headline ITT analysis, we will adopt the following approach for screening and addressing missing data. Screening stage: We will examine whether data is missing completely at random (MCAP), missing at random (MAR) or missing not at random (MNAR). A series of binary variables will be generated for all variables in the final ITT analysis that measures whether a case is missing (=1) or not (=0). Logistic regression will be used to examine whether missing data can be statistically accounted for using the other variables in the ITT analysis. When variables are found to account for a statistically significant proportion of variation in missing data, we will proceed to one of the next two stages. For instances where only data is missing in the teacher report SDQ externalising score outcome measure, we will add any additional covariates that were found from the screening stage to the final ITT model and re-estimate the effect size. For instances where data is missing in the baseline measure and where the screening stage identified variables that did account for variation in this missing data, we will construct a Multiple Imputation model using all variables listed for stage 1. The Multiple Imputation model will be estimated using 'STATA MI' to create 20 imputed data sets. These imputed data sets will be used to re-estimate the effect of Reach and the standard error (Rubin, 2004).

Compliance

The ITT analysis provides the most robust estimate of the *causal* impact of the Reach programme on the primary outcome. This is because the ITT analysis focuses on preserving randomisation and so best ensuring that the only difference between the Reach and the Control groups is group membership (all other differences are random). However, the ITT analysis does not take compliance with the Reach programme into account. In other words,

the ITT analysis captures the causal impact of Reach for YP who are randomised to the Reach programme. To estimate the impact of Reach for YP randomised to the Reach programme *and* who received the programme as intended (and specified), a compliance analysis will be undertaken. A combination of fidelity and dosage of Reach will be drawn on to construct a compliance variable at the YP level. This will identify a subsample of YP in the Reach intervention group who spent the required amount of time with their YW *in which* ALL of the core components of Reach were covered. This would be defined as receiving a minimum of 32 sessions with YW, and receiving all core components to be compliant. Detailed information on "time per core component" is being collected by the delivery team which will inform this variable. Associations between fidelity and dosage and outcomes will be undertaken as IPE analyses (see protocol for details on this) whilst the binary compliance variable (1 = compliant, 0 = not compliant) will be used to estimate the Compliers Average Causal Effect (CACE) using an instrumental variable and two stage least squares (2SLS) approach. The CACE analyses will provide the best estimate for the impact of Reach for YP who have received the programme as intended (and specified) in terms of fidelity and dosage. However, because CACE does not preserve randomisation, causal conclusions cannot be drawn from these analyses. The ITT and CACE analyses together provide two perspectives; an estimate of the causal impact of being allocated to receive Reach (ITT) and an estimate of the impact of receiving reach as intended and specified (CACE). This analysis will address a further impact evaluation research question as follows:

RQ15 What is the Compliance Average Causal Effect for the Reach programme on the primary outcome?

Intra-cluster correlations (ICCs)

For the primary outcome measure and corresponding baseline measure (teacher report SDQ externalising score), ICCs at the school level will be estimated using the 'estat icc' command in Stata. In the analysis section, a table will be included that presents the variance decomposition for the two levels, school and pupil, along with the ICC estimates.

Presentation of outcomes

Effect sizes will be calculated using Hedges' g, as specified in the following equation, where T is the treatment mean, C is the control mean, δ^2_{sch} is the school level variance and δ^2_{pup} is the pupil level variance for the null/empty model:

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{pup}}}$$

The headline effect size will be calculated from the group allocation (intervention/control) coefficient in the full analysis model, with the unconditional variance used as the denominator. The effect sizes will be reported along with confidence intervals and p-values to reflect statistical uncertainty.

Longitudinal follow-ups

No longitudinal follow-ups will be undertaken as part of the Reach Evaluation. However, the data will be submitted to the YEF data archive to enable long term follow up by others.

References

Austin, Peter (2008). Assessing balance in measured baseline covariates when using many-to-one matching on the propensity score. Pharmacoepidemiology and drug safety 17: 1270-1225.

Elff, M et al. (2021) *Multilevel Analysis with few clusters: Improving Likelihood-base methods to provide unbiased estimates and accurate inference.* British Journal of Political Science (2021), 51, 412–426

Hedges, L. V., and Hedberg, E. C. 2013. Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. Evaluation review, 37(6), pp. 445-489.

Hewitt, Catherine and David Togerson (2006). Is restricted randomisation necessary? British Medical Journal Research Methods 332: 1506-1508.

Ho, Daniel; Kosuke Imai; Gary King and Elizabeth Stuart (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis 15: 199-236.

Rubin, D. (2004). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Willis, B., Reaney-Wood, S., Demack, S., Jay, T., & Harris-Evans, J. (2023). Adventure Learning: Randomised Controlled Trial. Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Adventure-Learning-Final-Report.pdf?v=1688634116

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413