STATISTICAL ANALYSIS PLAN

Nurturing Empathy before Transition (NEBT) a split-cohort clustered randomised controlled efficacy trial

Sheffield Hallam University

Principal investigator: Sarah Reaney-Wood

YOUTH ENDOWMENT FUND

Project title

Statistical analysis plan

Evaluating institution: Sheffield Hallam University
Principal investigator(s): Dr Sarah Reaney-Wood

YEF statistical analysis plan

Project title[1]	*Nurturing Empathy before Transition (NEBT) a split-cohort clustered randomised controlled efficacy trial*
Developer (Institution)	*Roots of Empathy*
Evaluator (Institution)	*Sheffield Hallam University*
Principal investigator(s)	*Dr Sarah Reaney-Wood*
SAP author(s)	*Sarah Reaney-Wood, Sean Demack*
Trial design	*Split-cohort two-armed cluster randomised controlled trial with random allocation at the school level. Two Y5 pupil cohorts over 2 years (2022/23 & 2023/24).*
Trial type	Split Cohort (2 cohorts) Efficacy Trial
Evaluation setting	School
Target group	*Year 5 pupils*

[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).

Number of participants	***Planned allocation**: 4200 pupils in 140 schools across two years (60 schools in year one and 80 schools in year two). One class of 30 Y5 pupils per school.* ***Achieved**: 1662 pupils in 87 schools across two years (655 pupils in 33 schools in year 1; 1007 pupils in 54 schools in year 2). Mean number of pupils/school = 19.*
Primary outcome and data source	*Self-reported Me & My Feelings (M&MF) behavioural difficulties scale (Deighton et al, 2012)*
Secondary outcome and data source	*Self-reported M&MF emotional difficulties scale (Deighton et al, 2012)* *Self-reported Basic Empathy Scale (BES) affective empathy and cognitive empathy scales (Joliffe & Farrington, 2006).* *Teacher reported Strengths & Difficulties Questionnaire (SDQ); total difficulties and prosocial scales (Goodman, 2001).* *SDQ; hyperactivity, emotional problems, conduct problems and peer problems subscales (Goodman, 2001).*

SAP version history

Version	Date	Changes made and reason for revision
1.1	March 2024	Responding to peer review comments & changes to evaluation team.
1.0 [*original*]		*[leave blank for the original version]*

Any changes to the design or methods need to be discussed with the YEF Evaluation Manager and the developer team prior to any change(s) being finalised. Describe in the table above any agreed changes made to the evaluation design. Please ensure that these changes are also reflected in the SAP (CONSORT 3b, 6b).

Table of contents

Introduction

The impact of the Roots of Empathy Nurturing empathy before transition (RoE NEBT) programme will be estimated using a two-armed clustered randomised controlled trial (CRT). To meet the needs of RoE in terms of capacity, this efficacy trial uses a split cohort CRT design; cohort 1 running in 2022/23 and cohort 2 in 2023/24. The NEBT programme is designed to develop empathy and prosocial behaviour in participating pupils. In each cohort, year 5 pupils will participate in 27 45minute (approx.) sessions, three per month, delivered by a member of their school staff (usually a TA or SENCO) that has been trained as a RoE instructor. Through the sessions, each month a new theme will be covered, with nine themes in total:

Theme 1 – Meeting the Baby

Theme 2 – Crying

Theme 3 – Caring and Planning for Baby

Theme 4 – Relationships

Theme 5 – Sleep

Theme 6 – Safety

Theme 7 – Communicating

Theme 8 – Who am I?

Theme 9 – Goodbye and Good Wishes

Further information on the NEBT intervention can be found here: https://youthendowmentfund.org.uk/funding/who-we-fund/roots-of-empathy/

The aim of this trial is to explore what impact the NEBT programme has on pupils' emotional and behavioural difficulties, empathy and behaviour.

Design overview

The Nurturing Empathy Before Transition (NEBT) trial is an efficacy trial that is formed by combining data from two smaller trials that ran in 2022/23 and 2023/24 (a split-cohort design). This evaluation was commissioned prior to the Covid pandemic in 2019 but

postponed until 2021/22. Post-Covid recruitment difficulties led to the decision to split the efficacy trial over two years Both smaller evaluations adopted a two-armed design with randomisation at the school level. Prior to randomisation, schools identified a teacher and their Y5 class of pupils who would receive the NEBT programme should the school be randomised to the intervention group. Additionally, prior to randomisation, schools completed a baseline pupil survey (to collect M&MF and BES) and teacher survey (to collect teacher SDQ).

The aim for the first year was to recruit 60 schools for cohort 1 and 80 schools in cohort 2. Each school would have an estimated 30 pupils, resulting in a total sample of 4,200 pupils across 140 schools for the combined cohort.

The timetable for data collection across the two years can be found in table 2 below. Control schools will be operating business as usual during the trial and receive an incentive payment of £400 to recognise their commitment and effort towards participating in the trial.

Randomisation was conducted by the evaluation team at Sheffield Hallam University (SHU). Schools were randomised following baseline testing and prior to the programme starting. Randomisation was stratified by geographical area so that around half of schools in each area were randomised to the NEBT intervention or control groups. One difference between the two cohorts was the inclusion of primary schools in Wales for cohort 2. This was done following discussion between YEF, RoE and ourselves and pragmatically agreed to help maximise the recruited sample for cohort 2.

The primary outcome measure is the behavioural difficulties subscale from the Me and My Feelings (M&MF) questionnaire. The secondary outcomes are the emotional difficulties subscale of the M&MF questionnaire, the subscales of affective empathy and cognitive empathy from the Basic Empathy scale (BES) and the teacher Strengths and Difficulties questionnaire (SDQ) total difficulties and prosocial scales and the teacher SDQ emotional problems, conduct problems, peer problems and hyperactivity subscales. Data for all measures will be collected directly from pupils (M&MF and BES) and from teachers (SDQ).

Table 1 Summary of trial design

Trial design, including number of arms	*Split-cohort, two-armed, cluster randomised controlled trial*
Unit of randomisation	*School*

Stratification variables (if applicable)		*Geographic area*
Primary outcome	variable	*Behavioural difficulties at outcome*
	measure (instrument, scale, source)	*Behavioural difficulties as measured using the Self-report Me and My Feelings questionnaire behavioural difficulties sub scale (Deighton et al, 2012) [0 to 12 scale]*
Secondary outcome(s)	variable(s)	*Emotional difficulties, cognitive empathy, affective empathy, teacher reported behaviour difficulties, prosocial behaviour, hyperactivity, emotional problems, conduct problems and peer problems*
	measure(s) (instrument, scale, source)	*M&MF Emotional Difficulties subscale, [0 to 20 scale]* *BES cognitive empathy subscale, [9 to 45 scale]* *BES affective empathy subscale.[11 to 55 scale]* *Teacher SDQ total difficulties subscale [0 to 40 scale]* *Teacher SDQ prosocial score [0 to 10 scale]* *Teacher SDQ hyperactivity score [0 to 10 scale]* *Teacher SDQ emotional problems score [0 to 10 scale]* *Teacher SDQ conduct problems score [0 to 10 scale]* *Teacher SDQ peer problems score [0 to 10 scale]*
Baseline for primary outcome	**Variable**	*Behavioural difficulties at baseline*
	measure (instrument, scale, source)	*M&MF Behavioural difficulties subscale [0 to 12 scale]*
Baseline for secondary outcome	**Variable**	*Emotional difficulties, cognitive empathy, affective empathy, teacher reported behaviour difficulties & prosocial behaviour*
	measure (instrument, scale, source)	*M&MF Emotional Difficulties subscale, [0 to 20 scale]* *BES cognitive empathy subscale, [9 to 45 scale]* *BES affective empathy subscale.[11 to 55 scale]* *Teacher SDQ total difficulties subscale [0 to 40 scale]* *Teacher SDQ prosocial score [0 to 10 scale]* *Teacher SDQ hyperactivity score [0 to 10 scale]* *Teacher SDQ emotional problems score [0 to 10 scale]* *Teacher SDQ conduct problems score [0 to 10 scale]* *Teacher SDQ peer problems score [0 to 10 scale]*

Figure 1 illustrates the data collection activities at baseline and outcome and Table 2 provides more details.

Figure 1 Data collection activity for impact evaluation of the RoE NEBT programme

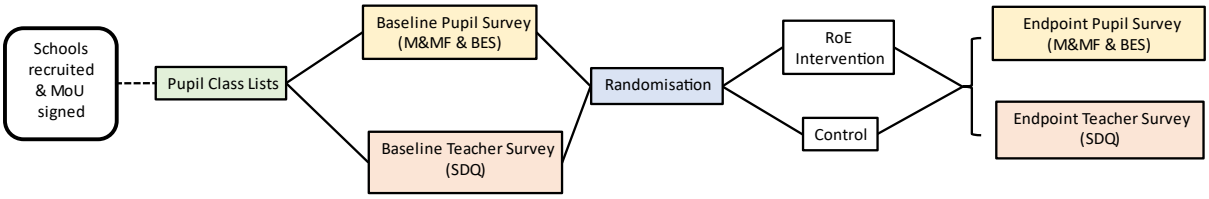


Table 2 Summary of data collection activity

Month/year	***Data Collection Activity***
Cohort 1	
September 2022	Cohort 1 schools that signed MoU supply pupil/teacher class lists for Y5 class taught by teacher selected to participate in RoE NEBT if their school was randomly selected.
Sept-Oct 2022	Baseline cohort 1 one schools **Pupil Survey:** Me & My Feelings (M&MF) and Basic Emapthy Scale (BES) **Teacher Survey:** Strengths and Difficulties questionnaire (SDQ)
June-July 2023	Immediate post-test cohort one schools **Pupil Survey:** Me & My Feelings (M&MF) and Basic Emapthy Scale (BES) **Teacher Survey:** Strengths and Difficulties questionnaire (SDQ)
Cohort 2	
September 2023	Cohort 2 schools that signed MoU supply pupil/teacher class lists for Y5 class taught by teacher selected to participate in RoE NEBT if their school was randomly selected.
Sept-Oct 2022	Baseline cohort two schools **Pupil Survey:** Me & My Feelings (M&MF) and Basic Emapthy Scale (BES) **Teacher Survey:** Strengths and Difficulties questionnaire (SDQ)

June-July 2024	Immediate post-test cohort two schools **Pupil Survey:** Me & My Feelings (M&MF) and Basic Emapthy Scale (BES) **Teacher Survey:** Strengths and Difficulties questionnaire (SDQ)

Sample size calculations overview

Please see Appendix 1 for a draft CONSORT flow diagrams for the primary outcome for cohorts 1, 2 and the combined efficacy sample. At the time of writing, data collection for cohort 1 is complete and the baseline data for cohort 2 has been collected. This data is drawn on to inform estimates for the primary outcome (M&MF Behavioural Difficulties) used for the power analyses presented below.

For cohort 1, 58 schools were recruited and had signed a Memorandum of Understanding (MoU) by Summer 2022. 33 of these schools supplied teacher/pupil details, participated in the baseline testing and were randomised in October 2022,

For cohort 2, 104 schools had signed an MoU by Summer 2023. 54 of these schools supplied teacher/pupil details and participated in the baseline testing and were randomised in October 2023.

In Cohort 1, 33 schools with 655 pupils were randomised to NEBT (16 schools, 331 pupils) or control (17 schools, 324 pupils). The primary outcome was collected for 275 pupils in 19 schools. This represents attrition of 58% at the pupil level and 42% at the school level.

In Cohort 2, 54 schools with 1007 pupils were randomised to NEBT (30 schools, 579 pupils) or control (24 schools, 428 pupils).

At the point of randomisation, there was a combined sample of 1662 pupils in 87 schools (910 pupils in 46 schools in the intervention group and 752 pupils in 41 schools in the control group). If attrition for cohort 1 is acknowledged, we can provide a range of estimates for the final (analysis) sample for a range of attrition assumptions for cohort 2:

The complete case sample for cohort 1 was 275 pupils in 19 schools.

- If 0% attrition is assumed for cohort 2, this would form a combined sample of 1282 pupils in 73 schools ~ an overall attrition of 23% at the pupil level and 16% at the school level.
- If 10% attrition (at pupil and school levels) is assumed for cohort 2, this would form a combined sample of 1181 pupils in 68 schools ~ an overall attrition of 29% at the pupil level and 22% at the school level.
- If 25% attrition is assumed for cohort 2, this would form a combined sample of 1030 pupils in 60 schools ~ an overall attrition of 38% at the pupil level and 32% at the school level.

- If 50% attrition is assumed for cohort 2, this would form a combined sample of 779 pupils in 46 ~ an overall attrition of 53% at the pupil level and 47% at the school level.

Please note, that even with 100% completion of outcome testing for Cohort 2, attrition at the pupil level for the primary outcome will be above 20%. Additionally, please see below for detail on the limitations for missing data analyses and imputation due to the limited pupil-level data (none other than the specified outcomes). This highlights the importance of response for the primary outcome and resource will be prioritised to maximise response for the cohort 2 pupil survey in June/July 2024.

MDES calculations were estimated using the formula set out by Bloom et al (2007) below and cross-checked using the Powerup! Software (Dong et al., 2015, sheet CRA2_2r).

$$MDES \sim \left(\frac{P}{1-P}\right) M_{(J-m-2)} \sqrt{\frac{ICC_2(1-R_C^2)}{J} + \frac{(1-ICC_2)(1-R_R^2)}{Jn}}$$

Where:

- P is the proportion of schools/clusters allocated to the intervention group set at 0.50 for protocol and 0.53 at randomisation stage (46 of the 87 schools).
- ICC_2 – Cluster (school) level Intra-cluster Correlation Coefficient (proportion of variance in the outcome between-schools). This was set at between 0.10 & 0.20 at the protocol stage and at 0.03 at randomisation stage drawing on cohort 1 for this estimate.
- R_C^2 is the covariate explanatory power at the cluster (school) level. This was set at 0.06 at the protocol stage and 0.58 at randomisation stage drawing on cohort 1 for this estimate.
- R_R^2 is the residual (within-school, between pupils) covariate explanatory power. This was set at 0.25 at the protocol stage and 0.27 at randomisation stage drawing on cohort 1 for this estimate.
- J is the total number of schools in the evaluation. This was set at 140 at the protocol stage and 87 at randomisation stage.
- n is the number of pupils-per-school. This was set at 30 at the protocol stage and 19 at randomisation stage.
- m is the number of cluster-level covariates included in the impact analyses. This is set at 7 (group membership, baseline measure at school & pupil levels, four dummy variables for the five geographical areas).
- M is the t-distribution multiplier and this has (J-m-2) degrees of freedom

Table 3 presents the minimum detectable effect size (MDES) estimates and sample sizes for the NEBT evaluation at protocol and randomisation stages.

A two-level clustered design was considered appropriate due to assumption that the NEBT intervention would be delivered in the same way across all geographical areas and the decision to stratify by geographical area was done to aid intervention delivery, rather than assuming differences in delivery between the areas. Geographical areas will be included in the analysis as school-level covariates (and are acknowledged in the power analyses).

No corrections for multiple testing have been made as the trial is a two-armed RCT, powered for the primary outcome only, with only one primary outcome for the pooled data for cohorts 1 and 2. As such, correction for multiple testing is not needed. Follow on sensitivity analyses will explore impact separately in cohorts 1 and 2.

Table 3 Sample and minimum detectable effect sizes (MDES)

Whole trial (year 1&2 combined)		**Protocol: Planned Sample Size at start of evaluation**	**Randomisation: Drawing on cohort 1 for ICC and correlation estimates.**	**Analysis**
Minimum Detectable Effect Size (MDES)		0.16-0.22	0.14	
Pre-test/ post-test correlations	level 1 (participant)	0.50 (R^2=0.25)	0.52	
	level 2 (cluster)	0.25	0.76	
Intracluster correlations (ICCs)	level 2 (cluster)	LOW=0.10 HIGH=0.20	0.03	
Alpha[2]		0.05	0.05	
Power		0.80	0.80	

Whole trial (year 1&2 combined)		Protocol: Planned Sample Size at start of evaluation	Randomisation: Drawing on cohort 1 for ICC and correlation estimates.	Analysis
One-sided or two-sided?		Two-sided	Two-sided	
Number of clusters[3]	intervention	70 schools	46 schools	
	control	70	41	
	total	140	87	
Number of participants	intervention	2100	910	
	control	2100	752	
	total	4200	1662	

With the lower ICC empirical estimate and higher than anticipated cluster (school) level correlation, at the point of randomisation, the sample of 1662 pupils in 87 schools across both cohorts would be sensitive to detect an effect size of 0.14 sds or higher as statistically significant ($p<0.05$, two tailed) with a statistical power of 0.80.

At the time of writing we have complete sample details for cohort 1 and baseline details for cohort 2 (see appendix). From this we can provide indicative MDES estimates for the final impact analyses of the M&MF Behavioural Difficulties primary outcome.

In cohort 1, attrition was notable (58% at pupil level and 42% at the school level) resulting in a (baseline & outcome) complete case sample of 275 pupils in 19 schools.

- If 0% attrition is assumed for cohort 2, this would form a combined sample of 1282 pupils in 73 schools which results in an indicative MDES estimate of 0.15 sds. [p=0.52]
- If 10% attrition (at pupil and school levels) is assumed for cohort 2, this would form a combined sample of 1181 pupils in 68 schools which results in an indicative MDES estimate of 0.16 sds.

[3] Please adjust as necessary e.g., for trials that are randomised at the setting, practitioner or participant level.

- If 25% attrition (at pupil and school levels) is assumed for cohort 2, this would form a combined sample of 1030 pupils in 60 schools which results in an indicative MDES estimate of 0.17 sds.
- If 50% attrition (at pupil and school levels) is assumed for cohort 2, this would form a combined sample of 778 pupils in 46 schools which results in an indicative MDES estimate of 0.20 sds.

Caution is needed in interpreting these indicative MDES estimates. Given that it is already known that cohort 1 had very high attrition at both pupil and school levels, randomisation is likely to have been undermined which in turn will weaken the validity from drawing causal conclusions from the impact analyses. MDES estimates provide an indication of statistical sensitivity for a (clustered) RCT of a specified size (and estimated parameters). Specifically, the MDES estimates assume that the only difference between the intervention and control groups is their group membership (one will experience the RoE NEBT programme, the other will not). All other differences are assumed to be random. This assumption is reasonable at the point of randomisation and when there is little/no attrition between randomisation and outcome data collection. However, this assumption becomes weaker with increasing attrition. This is because something other than randomness is likely to determine response (or lack of). For this reason, whilst the indicative MDES estimates suggest that the design is notably robust, they need to be interpreted with caution.

The very low school-level ICC estimate for M&MF Behavioural Difficulties outcome for cohort 1 along with the sizable explanatory power of the baseline M&MF (particularly at the school level) has resulted in a highly sensitive clustered RCT design. Unfortunately, this sensitivity needs to be considered alongside the sizable attrition for cohort 1. We will not have final details (on ICC or explanatory power) until we undertake the impact analyses for the combined cohorts. However, from cohort 1, the indication is that the initial size of this efficacy trial was unnecessarily large. A smaller trial provides good sensitivity and would bring advantages including cost, deliverer capacity and data collection activities. We will reflect on this again in the final report with the complete ITT sample for the combined cohorts.

Analysis

An intention-to-treat (ITT) approach will be used for all analyses. This means that regardless of whether a pupil continues with the trial they will be included (where possible) in the analysis. We will make attempts to collect the data from young people and schools that have withdrawn, but acknowledge that where schools/individuals have withdrawn it may be difficult to collect outcome data.

As the NEBT trial is taking place over two years (or cohorts), data from both cohorts will be combined for the main impact analysis. Sensitivity analysis detailed below will explore impact in the two cohorts to help assess the appropriateness of combining data from both cohorts.

Multi-level regression models will be conducted with pupils clustered within schools. Identifiers within the models are fixed effects based on the guidance provided in the YEF statistical analysis guide. The NEBT 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.

As noted in the 2024 protocol update, changes to the primary outcome resulted in redrafting the primary research question to focus on M&MF behavioural difficulties and a new research question (RQ2) has been added that focuses on M&MF emotional difficulties.

The following research questions will be answered by the impact evaluation:

Primary research question

1. RQ1: (Impact evaluation-primary outcome) What is the impact of the Nurturing Empathy programme on self-reported behavioural difficulties of primary school aged children when compared to a 'business as usual' control?

Secondary research questions

2. RQ2: What is the impact of the Nurturing Empathy programme on self-reported emotional difficulties of primary school aged children when compared to a 'business as usual' control?

3. RQ3: What is the impact of the Nurturing Empathy programme on self-reported affective empathy of primary school aged children when compared to a 'business as usual' control? (secondary outcome)

4. RQ4: What is the impact of the Nurturing Empathy programme on self-reported cognitive empathy of primary school aged children when compared to a 'business as usual' control? (secondary outcome)

5. RQ5: What is the impact of the Nurturing Empathy programme on teacher-reported prosocial behaviour of primary school aged children when compared to a 'business as usual' control? (secondary outcome)

6. RQ6: What is the impact of the Nurturing Empathy programme on teacher-reported school behaviour of primary school aged children when compared to a 'business as usual' control? (secondary outcome)

Exploratory

7. What is the difference in teacher-reported emotional problems (SDQ Sub scale) between the intervention group, when compared to a 'business as usual' control?

8. What is the difference in conduct problems (SDQ Sub scale) between the intervention group, when compared to a 'business as usual' control?

9. What is the difference in peer relationship problems (SDQ Sub scale) between the intervention group, when compared to a 'business as usual' control?

10. What is the difference in hyperactivity (SDQ Sub scale) between the intervention group, when compared to a 'business as usual' control?

The impact of NEBT 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 for the empty/null model:

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

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 for the Hedges' g effect size.

As recommended by Hedges & Hedberg (2013), pupil-level baseline scores will be centred around the school mean, and school baseline scores will be centred to the grand (school level) mean.

Primary outcome analysis

The primary outcome for the NEBT trial is pupil behavioural difficulties at endpoint using the behavioural difficulties subscale score of the Me and My Feelings questionnaire (Deighton et al., 2012). The M&MF behavioural difficulties scale has been selected as the primary outcome because previous studies have shown a reduction in behavioural difficulties (aggression)

following the RoE programme (Schonert-Reichl et al. 2012). In addition, whilst previous studies have reported an increase in prosocial behaviour, this didn't feel like the most appropriate primary outcome as the only way of assessing this was indirectly, through teacher reports, because to the age of the pupils. Previous research into the impact of Roots of Empathy programme has been dependent on teacher reports and this has been cited as a limitation.

Use of the M&MF scale, specifically behavioural difficulties as the primary outcome keeps this evaluation comparable with other YEF evaluations (by using an age appropriate YEF core measure) and in line with YEF strategy to provide evidence to prevent youth crime, whilst also overcoming the methodological limitations of previous evaluations.

The M&MF questionnaire is a 16-item school-based measure of child mental health, suitable for children aged 8-11. More details on the M&MF questionnaire can be found in the protocol, page 17. The M&MF behavioural difficulties subscale is the primary outcomes and will be calculated by adding items 11-16 of the scale, as illustrated below.

M&MF Behavioural Difficulties Scale

Calculated from summing 6 of the 16 M&MF items

M&MF Statement	**Response / Coding**
I get very angry I lose my temper I hit out when I am angry I do things to hurt people I break things on pupose	Never=0, Sometimes=1, Always=2
I am calm [Reverse Coded]	Never=2, Sometimes=1, Always=0

The M&MF behavioural difficulties subscale ranges between 0 and 12 where a higher score indicates higher (self-reported) behavioural difficulties. Change over time (i.e. baseline to outcome) might be negative (indicating reduced behavioural difficulties), null (no change) and positive (indicating increased behavioural difficulties).

Further details on the cut-offs for the M&MF scale can be found here: https://www.corc.uk.net/outcome-experience-measures/me-and-my-feelings-mmf/

The primary analysis will answer RQ1; multi-level linear regression models will be constructed that acknowledge that pupils are clustered in schools. In each of these two models, the endpoint M&MF score will be the outcome variable with the trial arm (1=NEBT or 0=Control) as the independent variable and baseline M&MF and geographical location as covariates.

Table 4 Example analysis model RQ1

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*
Empty model			Endpoint Me and My feelings (M&MF) behavioural difficulties
ITT sample (RQ1)	Baseline M&MF behavioural difficulties score (centred around school mean)	Group (1=RoE school, 0=Control school) Mean school-level baseline behavioural difficulties score (centred around school level Grand mean) Geographical area	

Secondary outcome analysis

The secondary outcomes are the M&MF emotional difficulties subscale (RQ2), affective empathy (RQ3) and cognitive empathy (RQ4) subscales of the Basic Empathy Scale (BES) (Pupil self-report) and six scales from the Strengths and Difficulties Questionnaire (SDQ) (Teacher report); prosocial behaviour (RQ5), total difficulties (RQ6), emotional problems (RQ7), conduct problems (RQ8), peer problems (RQ9) and hyperactivity (RQ10).

Basic Empathy Scale (BES)

The BES (Jolliffe & Farrington, 2006) is a 20-item questionnaire that assesses cognitive and affective elements of empathy. Both affective empathy and cognitive empathy are included as secondary outcomes. Models will be conducted as detailed for the primary outcome

analysis above to answer RQs 3 and 4. Within these models baseline affective empathy score OR cognitive empathy score will be included as a covariate as appropriate. Table 5 illustrate the analysis to be undertaken for the BES affective empathy score (RQ3).

Table 5 Example analysis model RQ3

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*
Empty Model			Endpoint Basic Empathy Scale (BES) affective empathy subscale
ITT sample (RQ3)	Baseline BES affective empathy score (centred around school mean)	Group (1=RoE school, 0=Control school) Mean school-level BES affective empathy score (centred around school level grand mean) Geographical area	

The Strengths and Difficulties Questionnaire (SDQ)

The teacher SDQ is a 25-item behavioural screening questionnaire for 4-17year olds. The items on the SDQ ask about a range of attributes, some negative and some positive. The total difficulties score for the SDQ is calculated by combining responses to 20 items whilst the prosocial score is calculated by combining responses to the remaining five items. The total difficulties score can be unpacked into four SDQ subscales (conduct problems, emotional problems, peer problems and hyperactivity) each with a 0 to 10 scale.

The Strengths and Difficulties Questionnaire is being used by YEF across its projects to create consistency and comparability between different evaluations. Further information about the SDQ is available here: https://www.sdqinfo.org/ .

A similar approach for constructing models for the two teacher-SDQ outcomes will be taken as specified for the primary outcome, to answer RQs 5 and 6 and exploratory RQs 7 to 10. Within these models baseline total SDQ and baseline SDQ subscales will be used as covariates (as appropriate). The exploratory analyses of the SDQ total difficulties subscales will adopt the same approach. Table 6 illustrate the analysis to be undertaken for the SDQ total difficulties score.

Table 6 Example analysis model RQ6

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	
ITT sample			Endpoint teacher-SDQ total difficulties score
ITT sample (RQ6)	Baseline teacher-SDQ total difficulties score (centred around school mean)	Group (1=RoE schools, 0=Control schools) Baseline teacher-SDQ total difficulties score (centred around school level Grand mean) Geographical area	

Subgroup analyses

This trial was commissioned as part of the first funding round for YEF after their set-up. As such, requirements for subgroup analyses were not set out. Discussions were had between all parties (YEF, SHU and RoE) and it was decided as part of this trial that collection of additional information from schools would be kept to a minimum and that follow on analysis could be undertaken at a later stage as part of the YEF archiving process.

This means that no data on participant gender, ethnicity, FMS-status or pupil-level details other than the specified outcomes were collected.

Further analyses

As sensitivity analyses, we will replicate the analyses specified above for the primary and secondary outcomes for both cohorts 1 and 2. For all outcomes, this will be done descriptively and by replicating the multilevel analyses for the separate cohorts. If these subsample analyses reveal evidence of differential impact in cohorts 1 and 2 for the M&MF behavioural difficulties primary outcome, we will also explore this directly by including two additional terms to the model outlined in Table 4 above; as illustrated in Table 7 below.

Table 7 Sensitivity analysis for primary outcome

Analysis and Sample	*Level 1 (pupil) Covariates*	*Level 2 (school) Covariates*	*Outcome Variable*

ITT sample (RQ1)	Baseline M&MF behavioural difficulties score (centred around school mean)	Group (1=RoE school, 0=Control school) Mean school-level baseline behavioural difficulties score (centred around school level Grand mean) Geographical area	Endpoint Me and My feelings (M&MF) behavioural difficulties
ITT sample (RQ1)	Baseline M&MF behavioural difficulties score (centred around school mean)	Group (1=RoE school, 0=Control school) Cohorts one (=0) or two (=1). Interaction Group*Cohorts Mean school-level baseline behavioural difficulties score (centred around school level Grand mean) Geographical area	

The inclusion of the 'cohorts' main effects dummy variable will replicate the analyses shown in the subsample analyses (cohorts 1 and 2) whilst the interaction term would directly capture whether the impact of RoE NEBT differed between the two cohorts.

In a spilt cohort trial is important that there are similar levels of fidelity in each year and that there are no changes to the Theory of Change (ToC) outlined within the protocol. This will be explored from the data collected as part of the implementation and process evaluation (IP data) and the compliance data below.

Longitudinal follow-up analyses

No longitudinal follow-ups will be undertaken as part of the NEBT RCT evaluation. However, Unique Pupil Identifiers (UPNs) have been collected by the evaluation team to enable long-term follow-up by others as part of the YEF data archiving process.

Imbalance at baseline

Our examination of imbalance at baseline will focus on the specified pupil-level self-reported outcomes (M&MF & BES), pupil-level teacher-reported outcomes (SDQ) and school level

statistics (%FSM, Ofsted, geographical area etc). These analyses will provide an indication of imbalance at baseline following randomisation. We will also examine how these pupil-level outcomes and school-level details are associated with the M&MF behavioural difficulties primary outcome.

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

Randomisation took place following the completion of baseline data collection and the ITT sample consists of pupils listed in the Y5 class who responded to the baseline pupil survey and completed the M&MF behavioural difficulties items. This means that there are no missing baseline data for the ITT sample in cohorts 1 or 2.

Missing data will arise due to attrition between baseline and outcome in cohorts 1 and/or 2. As noted above, attrition was sizable for the primary outcome in cohort 1 (58% at the pupil level). Even in the unlikely event of zero attrition for cohort 2, attrition for the combined efficacy trial will be above 20%.

The baseline and ITT samples will be compared to help illustrate the impact of missing data for the primary outcome variable only, Me and My Feelings (M&MF) behavioural difficulties score. This will firstly be done descriptively by tabulating missing cases across the categories of variables included in the ITT analysis (M&MF baseline). Reasons for any missingness will be summarised and a multi-level logistic regression model (1=in ITT model; 0=not in ITT model) will examine whether missingness is associated with school/intervention and/or pupil-level covariates. School level covariates include; geographical location, %FSM, %EAL, OFSTED. Pupil level covariates include; baseline M&MF (behavioural & emotional difficulties scale), baseline teacher-SDQ (prosocial and total difficulties scales), and baseline BEC (cognitive and affective empathy scales).

We have no plans to impute missing data for the primary outcome. This is because we do not have any complete pupil-level data to draw on for this imputation. So, whilst the patterns of

missingness will be examined, this will be to inform the interpretation of the ITT impact analysis rather than a stage before imputing missing cases.

Compliance

Compliance to the NEBT programme was defined at two levels; school (two criteria), and pupil (one criteria), these will be brought together in a final binary compliance variable at the pupil level. This is summarised in Table 8 and described in more detail below.

Table 8 Compliance components

Component	Data	Maximum	Minimum	Notes
School-level	Number of topics delivered from the NEBT curriculum, recorded by instructors	9 topics (3 sessions for each topic)	8 topics (3 sessions for each topic)	Focus is on number of topics and not sessions to assess the amount of the curriculum covered
	Instructor training attendance records collected by RoE	4-days	4-days	Full attendance at each day of training is required
Pupil-level	Number of topics pupil attended sessions for taken from attendance register	9 topics (3 sessions for each topic)	8 topics (3 sessions for each topic)	As with the school level the focus is on number of topics and not sessions to assess the amount of the curriculum covered

Compliance at the school level:

Criteria one: School level compliance will be measured by the number of NEBT topics delivered across the 9-month period. Data collected will be used to create a binary variable that indicates school level compliance, or not. The minimum threshold for number of NEBT topics delivered is eight (out of the total of nine). School level compliance will be summarised descriptively.

Criteria two: Number of RoE NEBT instructor training sessions that the TA/staff member attends. This will be measured on a categorical scale from 0 to 4. All four sessions must have been attended to be considered as compliant. The number of training sessions will be used alongside pupil level compliance below and brought into a binary variable.

Compliance at the pupil level:

Pupil level compliance uses the same threshold as school level in that pupils have to have attended sessions for 8 out of the 9 topics to be considered as having complied. For cohort 1, only four of the 16 schools randomised to NEBT provided pupil attendance data.

For the final binary compliance variable, we will draw together the pupil level topic variable and the instructor training variable. If both of these have been met, then this individual will be considered to have 'complied' (1). When one of these has not been met, the pupil will not be considered as having complied (0).

Whilst the compliance variable draws on data at both the school and pupil level, it is being brought together into a pupil level variable only. As such, multilevel CACE is not required.

For this analysis we are assuming that the control group are 100% non-takers in that they will not experience the RoE programme. This is an appropriate assumption given that this is a cluster RCT, meaning control pupils are very unlikely to receive RoE NEBT.

Compliance is not expected to be close to 100%. Firstly the endogenous assumption with will checked, if this is significant then and CACE analysis with Instrumental Variable (IV) approach using Two Stage Least Square (2SLS) analysis will be conducted on the primary outcome (Me and My Feelings behavioural difficulties score at endpoint) only (Sussman & Hayward. 2010; Tilbrook et al. 2014). If compliance is found not to be endogenous, we would switch to undertaking the CACE analysis as specified by Jo et al. (2008) and Schochet and Chiang (2011) summarised in the equation below:

$$CACE\ estimate = \frac{\text{ITT estimate}}{\text{proportion of pupils identified as 'compliant'}}$$

The purpose of the CACE analysis is to estimate the impact of the RoE programme for those pupils that can be considered compliant.

Intra-cluster correlations (ICCs)

The pre-test for the NEBT ROE trial will be Me and My Feelings (M&MF) behavioural difficulties subscale score at baseline (start of year 5) and the post tests will be M&MF behavioural difficulties subscale score at endpoint (end of year 5). For the pre and post-test, 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 (including geographical area and cohort), 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.

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.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. 2007. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis, 29(1), pp. 30-59.

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., Martin, A., Patalay, P. & Wolpert, W. (2013). The Development of a School-Based Measure of Child Mental Health. Journal of Psychoeducational Assessment, 31:247. https://doi.org/10.1177/0734282912465570

Deighton, J., Humphrey, N., Belsky, J., Boehnke, J., Vostanis, P., & Patalay, P. (2018). Longitudinal pathways between mental health difficulties and academic performance during middle childhood and early adolescence. British Journal of Developmental Psychology, 36(1), 110-126.

Dong, N., Kelcey, B., Maynard, R. and Spybrook, J. 2015 PowerUp! Tool for power analysis.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. Journal of the American Academy of Child & Adolescent Psychiatry, 40(11), 1337-1345. https://doi.org/10.1097/00004583-200111000-00015

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.

Humphrey, N., & Panayiotou, M. (2022). Bounce Back: randomised trial of a brief, school-based group intervention for children with emergent mental health difficulties. European Child & Adolescent Psychiatry, 31(1), 205-210.

Jo, B., Asparouhov, T., Muthen, B. and Ialongo, N. (2008) 'Cluster randomized trials with treatment noncompliance'. Psychological Methods, 13(1), 1–18.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. Journal of adolescence, 29(4), 589-611.

Patalay, P., Deighton, J., Fonagy, P., Vostanis, P., & Wolpert, M. (2014). Clinical validity of the Me and My School questionnaire: a self-report mental health measure for children and adolescents. Child and Adolescent Psychiatry and Mental Health, 8(17), https://doi.org/10.1186/1753-2000-8-17.

Schochet, P.Z. and Chiang, H.S. (2011) 'Estimation and identification of the complier average causal effect parameter in education RCTs'. Journal of Educational and Behavioral Statistics, 36(3), 307–345.

Schonert-Reichl, K.A., Smith, V., Zaidman-Zait, A. & Hertzman, C. (2012) Promoting children's prosocial behaviors in school: Impact of the "Roots of Empathy" program on the social and emotional competence of school-aged children. School Mental Health 4 pp1-21. https://doi.org/10.1007/s12310-011-9064-7

Sussman JB, Hayward RA. (2010) An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. BMJ. 2010 May 4;340:c2073. https://doi.org/10.1136/bmj.c2073

Tilbrook HE, Hewitt CE, Aplin JD, Semlyen A, Trewhela A, Watt I, Torgerson DJ. (2014) Compliance effects in a randomised controlled trial of yoga for chronic low back pain: a methodological study. Physiotherapy. 2014 Sep;100(3):256-62. https://doi.org/10.1016/j.physio.2013.10.001

van Noorden TH, Haselager GJ, Cillessen AH, Bukowski WM. (2015) Empathy and involvement in bullying in children and adolescents: a systematic review. J Youth Adolesc. 2015 Mar;44(3):637-57. https://doi.org/10.1007/s10964-014-0135-6 . Epub 2014 Jun 4. PMID: 24894581.

APPENDIX 1: CONSORT flow diagrams for primary outcome

Figure A1: Consort flow diagram for primary outcome– Cohort 1 (2022/23)

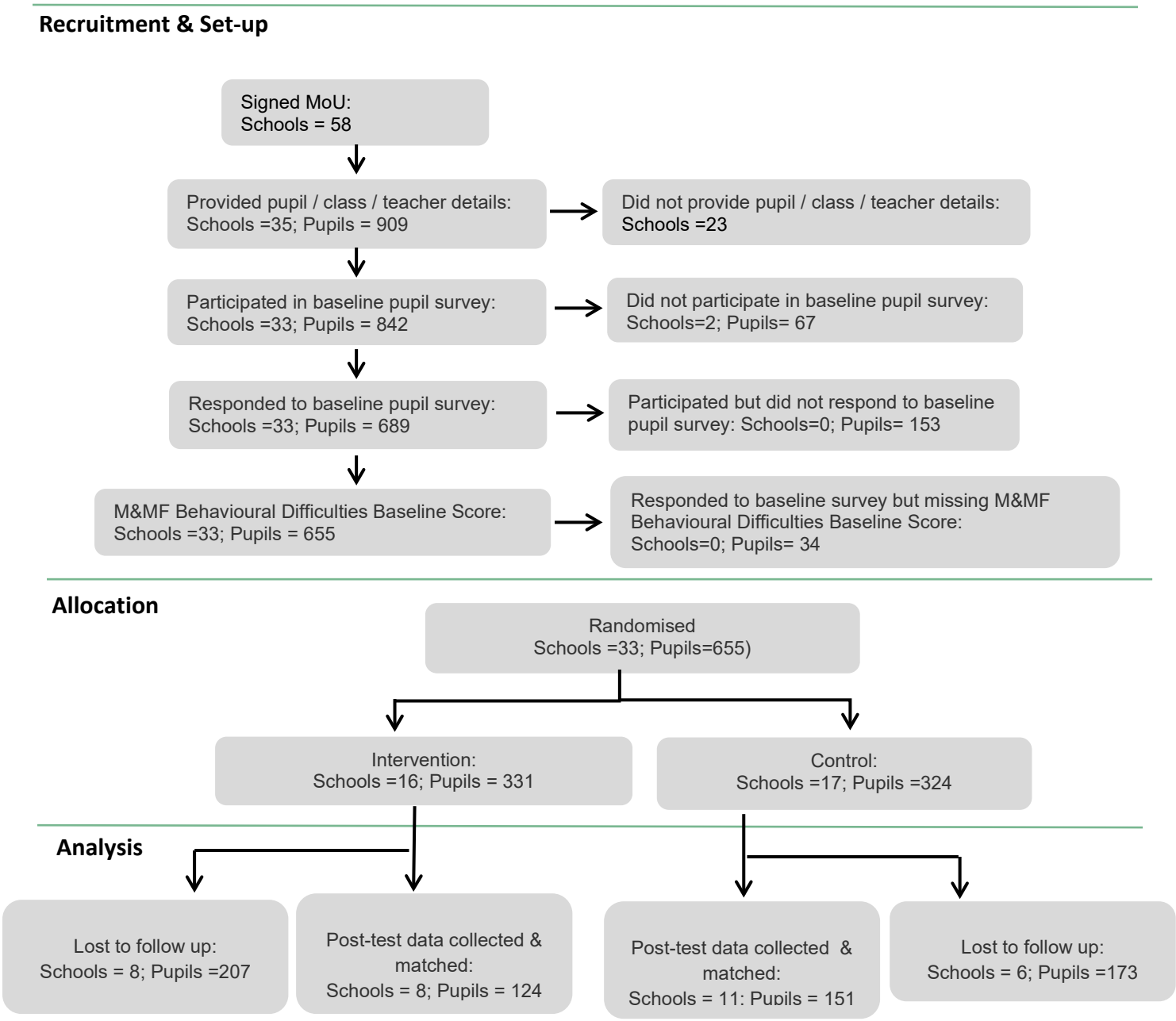


Figure A2: Consort flow diagram for primary outcome– Cohort 2 (2023/24)

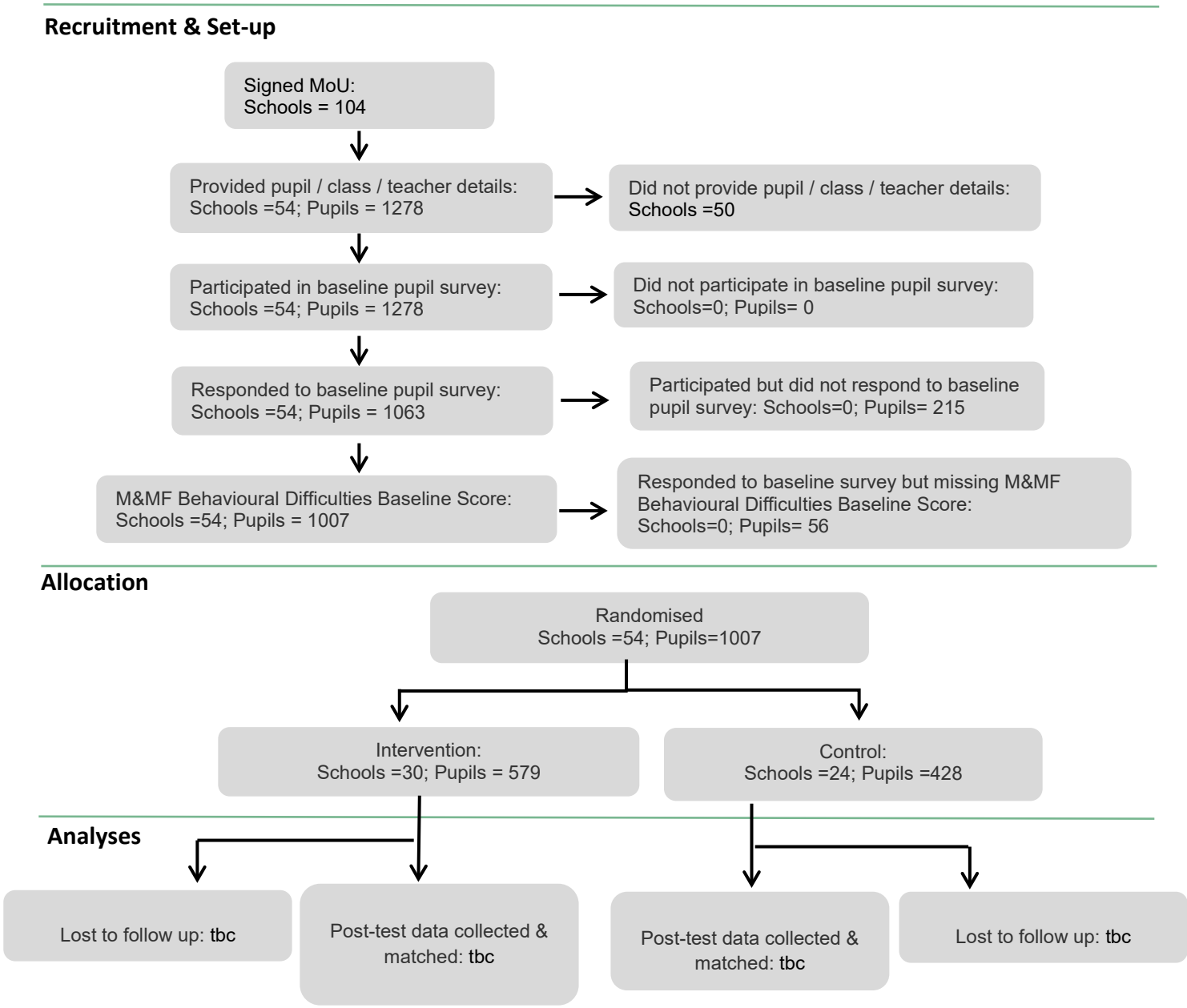
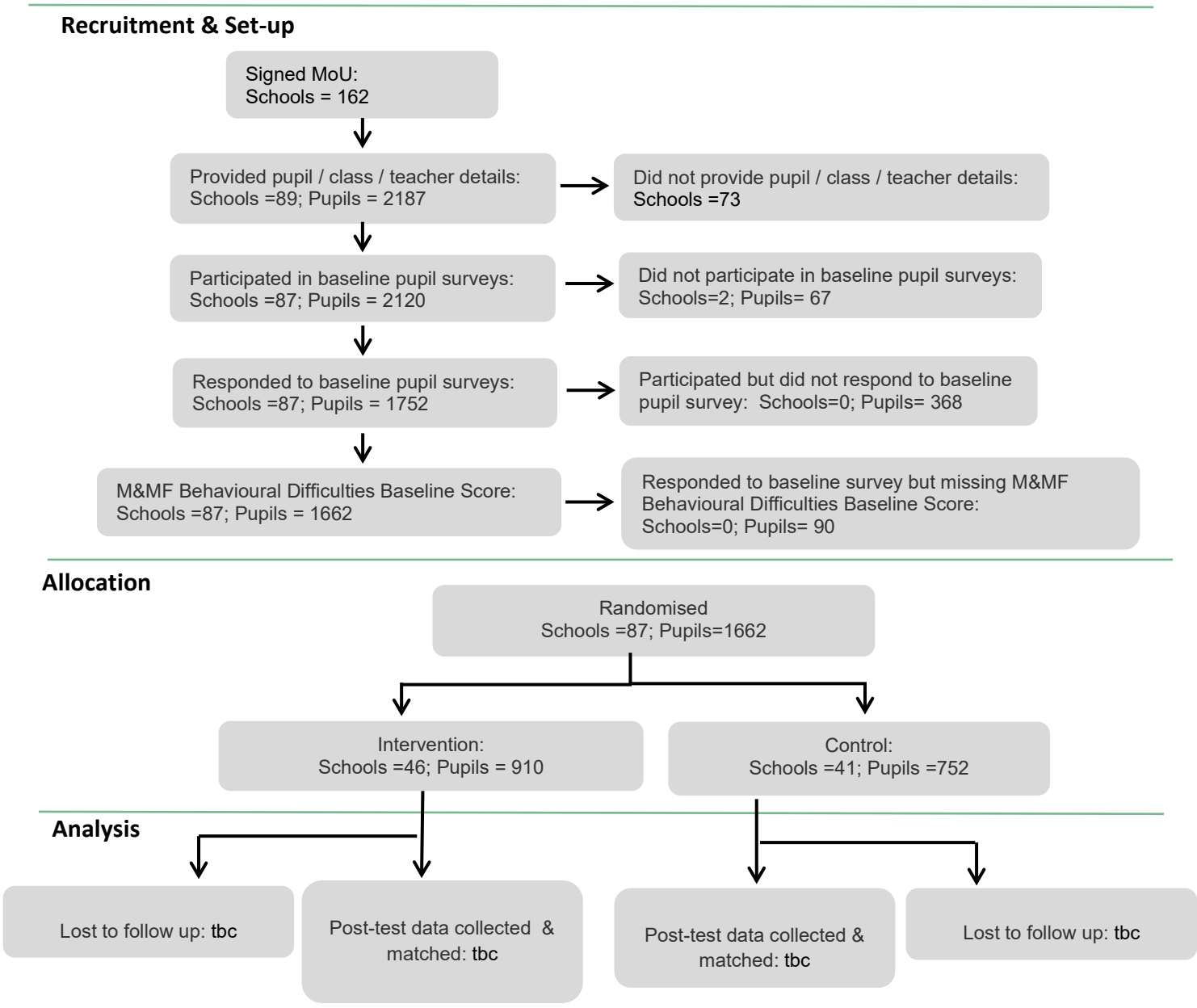


Figure A3: Consort flow diagram for primary outcome– Cohorts 1 & 2 combined



YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413