STATISTICAL ANALYSIS PLAN

Evaluation of the Inspiring Futures intervention: a cluster randomised controlled trial

Ipsos UK

Principal investigators: Dr Facundo Herrera, Prof. Stephen Morris, Jemuwem Eno-Amooquaye

YOUTH ENDOWMENT FUND

Evaluation of the Inspiring Futures intervention: a cluster randomised controlled trial

Statistical analysis plan

IPSOS UK

Principal investigators: Dr Facundo Herrera, Prof. Stephen Morris, Jemuwem Eno-Amooquaye

Project title[1]	Evaluation of the Inspiring Futures intervention: a cluster randomised controlled trial
Developer (Institution)	Rugby Football League (RFL)
Evaluator (Institution)	Ipsos UK
Principal investigator(s)	Dr Facundo Herrera, Prof. Stephen Morris, Jemuwem Eno-Amooquaye
SAP author(s)	Dr Facundo Herrera, Luisa Gomes
Trial design	Two-armed cluster randomised controlled trial with random allocation at the school level
Trial type	Efficacy
Evaluation setting	School
Target group	School pupils from Years 8 and 9
Number of participants	1,254 students across 114 school

[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).

Primary outcome and data source	Behavioural difficulties: SDQ – combined conduct and hyperactivity scales (0-20) - survey
Secondary outcome and data source	Internalising behaviour: SDQ – combined emotional problems and peer problems scales (0-20) - survey Pro-social behaviour: SDQ – Pro-social behaviour scale (0-10) - survey Children's wellbeing: Short Warwick Edinburgh Mental Wellbeing Scale – survey Number of temporary exclusions: Termly number - Administrative data Number of unauthorised absences: Termly number - Administrative data Number of authorised absences: Termly number - Administrative data Physical activity (self-reported 0-7) - survey

SAP version history

Version	Date	Changes made and reason for revision
1.0 [*original*]	31/03/25	

Table of contents

1 Introduction 4

2 Design overview 5

3 Sample size calculations overview 7

4 Analysis 10

4.1 Primary outcome analysis 13

4.2 Secondary outcome analysis 14

4.3 Subgroup analyses 14

4.4 Further analyses 15

4.5 Interim analyses and stopping rules 17

4.6 Longitudinal follow-up analyses 17

4.7 Imbalance at baseline 18

4.8 Missing data 21

4.9 Compliance 22

4.10 Intra-cluster correlations (ICCs) 23

4.11 Presentation of outcomes 23

5 References 24

1 Introduction

The intervention is called the Inspiring Futures Educate Mentoring Programme, which builds on the YEF pilot results that we published in 2023 (Wong et al., 2023). It is a cluster randomised controlled trial targeting young people recruited through schools. It targets young people aged 12-14 with an initial interest in sports and a record of poor behaviour and/or attendance. Mentoring programmes have shown to positively impact outcomes which are often associated with later involvement in violence (e.g., substance misuse, behavioural difficulties, educational outcomes, social connects, and emotional health) (Gaffney et al., 2023). Having a mentor can reduce the likelihood of offending by providing a positive role model.

The intervention consists of 12 weekly mentoring sessions over three months, delivered by the Rugby Football League, Foundation Delivery Partners, and delivery service Upshot. These sessions are delivered face-to-face, and they encompass personal wellbeing, collaboration and leadership. The aims of the sessions are:

- To build resilience, self-confidence and character in young people.
- To support positive choices and enable young people to engage positively with society.
- To improve critical thinking skills.
- To provide a healthy, stable, supportive framework at home and school.

The key mechanism of change is to use the sports element to encourage young people to develop an interest in rugby and a trusted relationship with their mentor. Having built a trusted relationship, mentors can provide emotional and social support to young people.

The intervention's outcomes are to address behavioural difficulties, including internalising and externalising behaviours and pro-social behaviour and wellbeing. Educational outcomes, such as attendance and attainment, are also expected to improve.

The programme will be evaluated through an efficacy, two-arm, cluster-randomised controlled trial (cRCT) with random allocation at the school level. Every school that signs up to participate in the trial will have an equal probability (50%) of being assigned to the treatment or control group. In addition, the trial is complemented by an Implementation and Process Evaluation (IPE).

The Implementation and Process Evaluation involves in-depth interviews with RFL coaches and case studies. The in-depth interviews aim to assess the extent to which the intervention is implemented as intended throughout the school, ensuring fidelity to the intervention's principles. The case studies will include in-depth interviews with staff members and students

and focus groups with students. The focus group aims to explore how the programme has been delivered within a particular school, including what has worked well or less well, staff's perception of the intervention, and factors affecting the implementation.

2 Design overview

This efficacy trial employs a two-arm cluster randomised controlled design with schools as the unit of randomisation. The trial targets students aged 12-14 (Years 8 and 9) within schools. Students screened for inclusion and aged 12-14 in schools allocated to the intervention form the intervention group and receive the rugby-based mentoring programme, whilst students aged 12-14 screened for inclusion in the programme but in schools randomised to control the control group and continue with business as usual.

The trial is being implemented in two waves:

- **Wave One**: Baseline data collection occurred in November-December 2024, with randomisation in December 2024. The intervention ran from January to mid-April 2025, with follow-up measurements scheduled for June 2025.
- **Wave Two**: Baseline data collection began in March 2025 and continues through April 2025, with randomisation planned for April 2025. The intervention will run from May to July 2025, with follow-up measurements from September to October 2025.

Randomisation was stratified by foundation only, rather than by both foundation and FSM6 as initially planned, to prevent empty stratification cells[2].

The primary outcome is externalising behaviour, measured using the Strengths and Difficulties Questionnaire (SDQ) combined conduct and hyperactivity scales. Secondary outcomes include internalising behaviour, pro-social behaviour, wellbeing, school exclusions, absences, and physical activity. These are measured through surveys and administrative data.

Eligible schools must have pupils in Years 8 and 9, and must not be fee-paying, alternative provision, special schools with a "SEMH" focus, or participating in another randomised trial. These criteria ensure generalisability and prevent contamination from other interventions.

[2] Empty cells would occur if all schools within a particular foundation fell into the same FSM6 category (either all above or all below the median). This was a realistic concern as some foundations had very few participating schools (as few as 2-4 schools). For example, if a foundation with only two schools had both schools above the FSM6 median, this would create an empty cell in the "below median" category for that foundation, potentially compromising the randomisation balance. The stratification approach is discussed further in the Imbalance at baseline section.

Table 1 summarises the trial design, including details of the outcome measures.

Table 1 Summary of trial design

Trial design, including number of arms		Two-arm, cluster-randomised
Unit of randomisation		Cluster (school)
Stratification variables (if applicable)		11 Foundations
Primary outcome	variable	Behavioural difficulties
	measure (instrument, scale, source)	Externalising behaviour: SDQ – combined conduct and hyperactivity scales (0-20) - survey
Secondary outcome(s)	variable(s)	Internalising behaviour Pro-social behaviour Children's wellbeing Number of temporary exclusions Number of unauthorised absences Number of authorised absences Amount of physical activity
	measure(s) (instrument, scale, source)	Internalising behaviour: SDQ – combined emotional problems and peer problems scales (0-20) - survey Pro-social behaviour: SDQ – Pro-social behaviour scale (0-10) - survey Children's wellbeing: Short Warwick Edinburgh Mental Wellbeing Scale – survey Number of temporary exclusions: Termly number of fixed-term exclusion events - Administrative data (NPD) Number of unauthorised absences: Termly number - Administrative data (NPD)

		Number of authorised absences: Termly number - Administrative data (NPD) Amount of physical activity: Self-reported question (0-7) - survey
Baseline for primary outcome	**variable**	Behavioural difficulties
	measure (instrument, scale, source)	Externalising behaviour: SDQ – combined conduct and hyperactivity scales (0-20) - survey
Baseline for secondary outcome	**variable**	Internalising behaviour Pro-social behaviour Children's wellbeing Number of temporary exclusions Number of unauthorised absences Number of authorised absences Amount of physical activity
	measure (instrument, scale, source)	Internalising behaviour: SDQ – combined emotional problems and peer problems scales (0-20) - survey Pro-social behaviour: SDQ – Pro-social behaviour scale (0-10) - survey Children's wellbeing: Short Warwick Edinburgh Mental Wellbeing Scale – survey Number of temporary exclusions: Termly number of fixed-term exclusion events - Administrative data (NPD) Number of unauthorised absences: Termly number - Administrative data (NPD) Number of authorised absences: Termly number - Administrative data (NPD) Amount of physical activity: Self-reported question (0-7) - survey

3 Sample size calculations overview

Table 2 summarises the sample size calculation, including the main parameters.

Sample size calculations were conducted using PowerUp software (Dong & Maynard, 2013), which implements optimal design procedures for multilevel randomised trials. Our calculations are based on a two-level random effects model accounting for the clustered nature of our data, with students (level 1) nested within schools (level 2).

The model assumes random assignment at the school level with baseline covariates included to improve precision. PowerUp's cluster randomised trial function was used to determine the minimum detectable effect size under our specified design parameters.

We determined our sample size requirements a priori through power calculations conducted during the co-design phase rather than being driven by practical constraints. Our sample size estimation aims to balance statistical robustness with practical implementation. The pilot study provided valuable insights on key parameters, including an intracluster correlation (ICC) of 0.10 for the Strengths and Difficulties Questionnaire (SDQ). However, our power calculations use more conservative assumptions based on analysis of comparable educational datasets by our statistical advisors.

We developed two scenarios using standard parameters of 0.05 alpha level, 80% power, and two-sided testing. Scenario 1 uses more conservative assumptions with pre-post correlation of 0.52 ($R^2 = 0.27$), closer to pilot study findings. Scenario 2 uses correlation of 0.63 ($R^2 = 0.40$) based on analysis of comparable datasets by our statistical advisors, representing improved measurement precision expected in the full trial. The higher correlation in Scenario 2 aligns with established SDQ test-retest reliability in school settings ranging from 0.40-0.81 (Stone et al., 2015; Turi et al., 2011).

Scenario 1 represents our primary scenario (102 schools, correlation 0.52) assuming 100% school retention to detect an effect size of 0.20. Scenario 2 provides a more conservative approach accounting for school attrition (114 schools, correlation 0.63), assuming 90% school retention. We adopted Scenario 2 to ensure adequate power despite potential attrition, achieving an MDES of 0.175.

The cluster design follows this structure: we aim to recruit 12 students per school across 114 schools, totalling 1,368 students. PowerUp calculations use the recruited sample of 1,254 students (accounting for some recruitment variation), with 73% expected retention built into the statistical model. This approach yields an MDES of 0.175 whilst accounting for both clustering effects and anticipated attrition.

Table 2 also shows the achieved MDES at baseline after randomisation pooling together all waves. The combined waves column presents the achieved power calculations based on our actual baseline sample of 1,133 participants across 99 schools. The observed intracluster

correlation for our primary outcome (0.18) is higher than assumed (0.10), reducing statistical power and resulting in an MDES of 0.221. However, ICCs for secondary outcomes remain far below 0.10, suggesting more favourable power for these analyses. This calculation maintains the assumed pre-post correlation of 0.63, which is a conservative estimate and remains to be confirmed at follow-up. The final statistical power will depend on both the actual pre-post correlation achieved and the eventual follow-up sample size.

Table 2 Sample size calculations

		Protocol Scenario 1	Protocol Scenario 2	Randomisation Combined waves
Minimum Detectable Effect Size (MDES)		0.185	0.175	0.221
Pre-test/ post-test correlations	level 1 (participant)	0.52	0.63	0.63[3]
	level 2 (cluster)	0.52	0.63	0.63[4]
Intracluster correlations (ICCs)	level 1 (participant)	n/a	n/a	n/a
	level 2 (cluster)	0.1	0.1	0.18
Alpha[5]		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two
Average cluster size		11	11	11.4
	intervention	51	57	52

[3] This is still an assumption

[4] Same as above

[5] Please adjust as necessary for trials with multiple primary outcomes, 3-arm trials etc. when a Bonferroni correction is used to account for family-wise errors.

		Protocol Scenario 1	Protocol Scenario 2	Randomisation Combined waves
Number of clusters[6]	control	51	57	49
	total	102	114	99
Number of participants	intervention	561	627	568
	control	561	627	565
	total	1,112	1,254	1,133

4 Analysis

The analytical approach was determined prior to baseline data collection. Stata© syntax is provided at the conclusion of this section for transparency. We will employ an intention-to-treat (ITT) framework, analysing all available data while maintaining participants in their originally assigned groups. Our analysis will address the research questions below.

Primary research question

- ERQ1: What is the mean difference in **externalising behaviour**, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomains of Conduct Problems and Hyperactivity, between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

Secondary research questions

- ERQ2: What is the mean difference in **internalising behaviours**, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomains of Emotional Problems and Peer Problems, between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

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

- ERQ3: What is the mean difference in **pro-social behaviours**, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomain of Pro-social behaviour, between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

- ERQ4: What is the mean difference in **wellbeing**, measured by the Short Warwick Edinburgh Wellbeing Scale (SWEMWBS), between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

- ERQ5: What is the mean difference in the percentage of **temporary exclusions** before/during/after the intervention term between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

- ERQ6: What is the mean difference in the percentage of **authorised absences** before/during/after the intervention term between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

- ERQ7: What is the mean difference in the percentage of **unauthorised absences** before/during/after the intervention term between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

- ERQ8: What is the difference in the number of days Children and Young People (CYP) have **engaged in physical activity for at least 30 minutes**, sufficient to elevate breathing rate, between CYP in intervention settings receiving RFL mentoring and CYP in control settings receiving business-as-usual at follow-up?

As evidence suggests, the distribution of the Conduct Problems and Hyperactivity subscale of the SDQ is normally distributed (Caldwell et al., 2021); thus, the primary analysis will take the form of a multilevel model with random effects at the school level, considering pupils are nested within schools and this may introduce variation. The model will include the binary treatment variable and be adjusted for baseline stratification covariates and the baseline value of the outcome.

We are employing a random-effects approach to model the cluster-level effects at the school level in our trial. This approach allows us to treat schools as random samples from a broader population, enabling the generalizability of our findings beyond just the sampled schools. Moreover, the random-effects model incorporates partial pooling or shrinkage, which can

lead to better predictions of school-level effects compared to a fixed-effects approach. Crucially, with a sufficient number of school clusters (more than 10 or 20) in our sample, we can reliably estimate the between-cluster variance and make valid inferences about the variability of school-level effects, providing insights into the impact of school-level factors on the outcome of interest (Rabe-Hesketh & Skrondal, 2008).

All outcomes are analysed at the individual level using multilevel modelling. Table 3 sets out the statistical analysis for each outcome by level.

Table 3 Statistical analysis by outcome

Outcome	Model	Covariates
Primary outcome Behavioural difficulties	Multilevel model	Pre-treatment scores of outcomes Demographic factors (sex, age)
Secondary outcomes Internalising behaviour Pro-social behaviour Children's wellbeing Number of temporary exclusions Number of unauthorised absences Number of authorised absences Amount of physical activity	Multilevel model	Pre-treatment scores of outcomes Demographic factors (sex, age)

The syntax in Stata© for a fully adjusted model would be as below, using the 'mixed' command:

mixed post-treatment_outcome RFL baseline_outcome sex age wave foundation|| school_id, reml

4.1 Primary outcome analysis

The primary outcome combines the Conduct Problems and Hyperactivity sub-scales of the Strengths and Difficulties Questionnaire (SDQ) to measure externalising behaviours.

We will use a two-level multilevel model to account for clustering in our data, where pupils are nested within schools. This approach treats participating schools as a random sample from the broader school population. Multilevel models effectively handle hierarchical data structures and properly account for variation both within schools (among pupils) and between schools. These models capture complex sources of variation across multiple levels of the hierarchy (Bosker & Snijders, 2011).

The two-level random-intercept model is described by:

$$Y_{ij} = \beta_0 + \beta_1 RFL_j + \beta_2 Baseline_{ij} + \beta_3 Wave_j + \beta_4 Strat_j + \mu_j + \varepsilon_{ij}$$

- Y_{ij} is the outcome for pupil i in school j
- RFL_j is a binary variable indicating intervention (1) or control (0) assignment
- $Baseline_{ij}$ represents pupil-level pre-test covariates
- $Wave_j$ is a binary revealing whether the school was in the first or second wave of the trial
- $Strat_j$ is a variable capturing the strata within which schools were randomised, namely, foundations
- μ_j are the school-level residuals [uj ~ N(0, σ^2u)]
- ε_{ij} are the individual-level residuals [eij ~ N(0, σ^2e)]

This model includes a random intercept where μ_j corresponds to the school-level intercept for school j. The total variance splits into between-school variance and within-school variance. The intervention effect is captured by β_1.

In this trial, we specified one primary and several secondary outcomes. While adjusting for multiple testing can reduce type I error risk (false positives), it increases type II error risk (missed effects). Given our clear distinction between primary and secondary outcomes, strict adjustment for multiple testing may be unnecessary. We will discuss our findings carefully in the final report, considering the context of multiple outcomes and balancing type I and type II error risks. This approach recognises that avoiding false positives should not prevent identifying genuine effects (Zhang et al., 1997).

4.2 Secondary outcome analysis

The remaining secondary outcomes at the pupil levels follow the same equation as above.

Absence and exclusion outcomes will be measured at termly level to align with the 12-week intervention period. We will compare treatment and control groups before, during and after the intervention term.

4.3 Subgroup analyses

Our subgroup analysis follows a pre-specified approach established in the protocol before data collection began. We will explore differential effects across three key subgroups: sex, ethnicity, and FSM status. Sex will be coded as a binary variable (male/female). FSM status will be binary (eligible/not eligible). For ethnicity, we will present effect estimates and confidence intervals for all ethnic groups with sufficient sample sizes ($n \geq 30$), recognising that smaller groups will have limited power for detecting effects. We will avoid artificial binary categorisations of ethnicity that may not reflect meaningful distinctions in our sample.

We acknowledge that subgroup analyses will be underpowered compared to our main analysis. These analyses are exploratory and will require cautious interpretation, particularly for smaller subgroups. Effect size estimates will be presented with wide confidence intervals, and we will emphasise the preliminary nature of any subgroup findings.

For each subgroup, we will employ two complementary analytical strategies. First, we will run separate multilevel regression models for each subgroup (e.g., separate models for boys and girls). This approach provides straightforward estimates of intervention effects within each group but does not directly test for differences between groups.

Second, we will run multilevel models with interaction terms between the treatment indicator and subgroup variables. This strategy formally tests whether intervention effects differ significantly between subgroups (e.g., whether the programme works differently for boys versus girls).

Additionally, we will conduct exploratory latent class analysis using baseline SDQ subscale scores (conduct problems, hyperactivity, emotional symptoms, peer problems, and prosocial behaviour) from all participants to identify naturally occurring behavioural profiles. We will apply this classification to analyse intervention effects on our primary outcome (behavioural difficulties) and key secondary outcomes (internalising behaviour and wellbeing), as these align most closely with the baseline behavioural patterns used to create the classes.

This approach is theoretically motivated by research suggesting that mentoring interventions may work through different mechanisms for children with different baseline risk profiles. For

example, children with predominantly externalising difficulties may benefit more from the programme's focus on self-regulation and leadership skills, whilst those with internalising problems may respond better to the peer support and confidence-building elements. This analysis will help identify which young people are most likely to benefit from rugby-based mentoring programmes.

4.4 Further analyses

To thoroughly assess intervention effects, we will conduct a series of multilevel regression analyses, progressively building complexity:

Model 1: Null model

We will begin with an empty multilevel model that includes only the outcome variable with random effects at the school level. This establishes the baseline variance components and intraclass correlation, showing how much outcome variation exists between schools before accounting for any explanatory variables.

Model 2: Treatment dummy only

The second model will add only the treatment indicator as a fixed effect while maintaining the multilevel structure. This provides an unadjusted estimate of the average treatment effect across all participants.

Model 3: Fully specified model

Our third model will include the treatment indicator alongside all pre-specified covariates (including baseline outcome measures, stratification variables, and relevant demographic characteristics). This represents our primary analysis model, providing adjusted estimates of intervention effects.

Model 4: Interaction model

We will extend the fully specified model by adding interaction terms between treatment and Free School Meal (FSM) status at the school level. This will formally test whether intervention effects differ for pupils from economically disadvantaged backgrounds. FSM at the school

level is captured by a dummy variable that indicates whether a school is above the median proportion of pupils with FSM6[7].

Exploratory subscale analyses

We will conduct exploratory analyses examining individual SDQ subscales to enhance comparability with other YEF evaluations and support future meta-analyses. These analyses will examine:

- Conduct problems (0-10 scale)
- Hyperactivity (0-10 scale)
- Emotional problems (0-10 scale)
- Peer problems (0-10 scale)

These subscale analyses will use the fully specified model (Model 3) to provide adjusted estimates. Results will be presented in appendices as supplementary findings to support cross-study comparisons whilst maintaining focus on our pre-specified combined scale outcomes.

Dosage analysis

Similarly to FSM, we will conduct a dosage analysis to examine how intervention impact varies according to treatment intensity received. This analysis will incorporate interaction terms between the treatment indicator and a continuous variable measuring the number of sessions attended (0-12 sessions).

This approach allows us to investigate whether a dose-response relationship exists between intervention exposure and outcomes. The continuous specification provides greater statistical power and avoids arbitrary categorisation of engagement levels. We will examine whether effects increase linearly with attendance or whether threshold effects exist at particular attendance levels.

Findings from this analysis will complement the Implementation and Process Evaluation by linking quantitative measures of implementation fidelity with outcome measures, providing insights into the minimum effective dose and optimal engagement levels for future programme delivery.

[7] FSM6 refers to pupils who have been eligible for free school meals (FSM) at any point in the preceding six years, and this is a key factor in how Pupil Premium funding is allocated to schools.

We will conduct the full four-model sequence for our primary outcome (externalising behaviour). For secondary outcomes, we will present the fully specified model (Model 3) only, as summarised in Table 4. This focused approach balances analytical rigour with clear, interpretable reporting.

Table 4 Summary of analysis

Outcome type	Model 1 (Null)	Model 2 (Treatment only)	Model 3 (Fully specified)	Model 4 (Interactions)
Primary (Externalising behaviour)	✓	✓	✓	✓ (FSM + Dosage)
Secondary outcomes			✓	

Findings from this analysis will complement the Implementation and Process Evaluation by linking quantitative measures of implementation fidelity with outcome measures.

4.5 Interim analyses and stopping rules

There is no interim analysis or stopping rules.

4.6 Longitudinal follow-up analyses

Our follow-up data collection will occur in two distinct waves to accommodate the staggered implementation of the intervention. The first wave of follow-up data will be collected in May 2025, capturing outcomes from schools that received the intervention in the initial delivery phase. The second wave will follow in September 2025, gathering data from schools in the latter implementation phase.

We will conduct separate analyses using this administrative data for our secondary outcomes that use National Pupil Database (NPD) measures. The NPD data collection runs parallel to our primary data collection but follows its distinct timeline.

The comprehensive analysis and triangulation stage will occur between December 2025 and March 2026. This timeline allows sufficient time for:

1. Processing and cleaning all primary data from both waves
2. Analysing the NPD administrative data separately
3. Running our full suite of analyses across all outcome measures
4. Triangulating findings across different data sources and models

4.7 Imbalance at baseline

The randomisation was conducted by foundation, which was our key stratification variable. An independent Ipsos team implemented the randomisation using Stata's 'randtreat' command. This approach efficiently addresses the "misfits" problem that occurs when schools cannot be distributed evenly across treatment arms within each foundation. The command maintains overall treatment balance whilst handling unequal allocations across strata where exact proportional division is not possible (Carril, 2017). Table 5 shows school allocation by foundation, demonstrating successful stratified randomisation with minor expected variation in smaller foundations. Post-randomisation attrition was minimal (2.9%) and balanced, maintaining the integrity of random allocation.

Table 5 Number of randomised schools by foundation and trial arm

Foundation	Control	Intervention	Total
Barrow Raiders	3	3	6
Huddersfield Giants	2	3	5
Hull FC	6	6	12
Leeds Rhinos	9	8	17
Leigh Leopards	7	7	14
Salford Red Devils	3	5*	8
St Helens	4*	5	9

Swinton Lions	1	3	4
Wakefield Trinity	5	5*	10
Warrington Wolves	6	5	11
Wigan Warriors	3	3	6
Total	**49**	**53**	**102**

Note: three schools withdrew post-randomisation: one control (St Helens) and two intervention (Salford Red Devils, Wakefield Trinity).

During the co-design phase, we initially considered stratifying by foundation and the proportion of pupils eligible for Free School Meals (specifically the FSM6 measure). However, after careful consideration, we decided to stratify by foundation only.

This decision was made for two primary reasons:

- Practical implementation considerations: Some foundations had a relatively small number of schools (as few as 2 or 4), which could have created empty cells in the stratification matrix if we had included FSM6 as an additional stratifying variable.
- Statistical stability: When multiple stratification variables are used with small numbers of units in some combinations, this can lead to imbalanced randomisation blocks, potentially reducing rather than improving the balance across treatment arms.

While FSM6 was not used as a stratification variable in the randomisation process, it will be included as a covariate in one of our analysis models to explore sensitivity to any potential imbalance between treatment and control groups on this important characteristic.

This approach follows best practice in RCT design, which advises limiting stratification variables, particularly with smaller sample sizes. Restricting stratification when working with few schools ensures randomisation remains practical and maintains statistical integrity (Donner & Klar, 2004).

Table 6 shows baseline balance between control and intervention groups for key survey measures (combined Wave 1 and Wave 2 data). Effect sizes are small across all measures: externalising behaviour (-0.14), internalising behaviour (-0.005), prosocial behaviour (0.05), and wellbeing (0.02). These negligible differences demonstrate that randomisation

successfully created comparable groups at baseline across our primary and secondary outcome measures.

Table 6 Baseline outcome measures by treatment allocation

Pupil-level (continuous)	Control group		Intervention group		Effect size (Hedges's g)
	N (missing)	Mean (SD)	N (missing)	Mean (SD)	
SDQ – Externalising behaviour (primary outcome)	565 (129)	10.30 (4.10)	568 (108)	10.88 (3.94)	-0.14
SDQ – Internalising behaviour	565 (129)	5.10 (3.39)	568 (108)	5.10 (3.37)	-0.005
SDQ – Prosocial	565 (129)	6.70 (2.15)	568 (108)	6.59 (2.0)	0.05
Short Warwick Edinburgh Mental Wellbeing Scale	557 (137)	22.13 (5.02)	561 (115)	22.01 (5.10)	0.02

Table 7 Demographic variables at baseline

Categorical variable	Control group		Intervention group	
	N(missing)	%	N(missing)	%
BAME population	544 (138)	16.73%	547 (116)	16.10%

Continuous variable	N(missing)	Mean	N(missing)	Mean
Age	579 (115)	12.98	592 (84)	12.98

4.8 Missing data

We will only analyse missing data patterns for the primary outcome (externalising behaviour measured five months post-randomisation). This analysis will document missing data proportions and identify systematic patterns in the dataset.

Some participant attrition is expected in educational trials. Our sample size calculations accounted for this, assuming 27% participant- and 10% school-level attrition. To understand the missing data mechanism, we will investigate whether or not data is missing completely at random or at random. We will compare baseline characteristics for each trial arm between participants with complete follow-up data and those lost to follow-up. A logistic regression model will help identify any systematic differences between groups. We can reasonably consider the data missing randomly if this model reveals no significant predictors.

To maintain statistical power, our approach to handling missing outcome values will depend on the proportion missing:

- Below 5%: We will exclude these observations as they have minimal impact on the results
- Between 5-40%: We will implement multiple imputation techniques
- Above 40%: Multiple imputation becomes less reliable; we will acknowledge this limitation in our interpretation

Before imputation, we will conduct a variable reduction analysis examining associations between baseline variables and missing follow-up data. The imputation model will include only variables with meaningful associations (p-value < 0.10).

The number of imputations will be determined to achieve at least 96% statistical efficiency, calculated based on the fraction of missing values and required repetitions.

Our multiple imputation procedures will apply the same statistical model and assumptions used in the primary outcome analysis. If evidence suggests data is missing not at random, or

if missing data patterns correlate with trial allocation, we will conduct sensitivity analyses using pattern mixture models (Hedeker & Gibbons, 1997; Little, 1993). In practice, this approach involves creating separate analytical models for different groups based on their missingness patterns - for example, one model for participants who completed all assessments, another for those who dropped out after baseline, and a third for those who missed only the follow-up assessment. Each model makes different assumptions about why data is missing and what the unobserved outcomes might have been. By comparing results across these different models, we can assess whether our conclusions change depending on these assumptions, thereby testing the robustness of our findings to different missing data scenarios.

4.9 Compliance

School-level compliance will be assessed using two dimensions from the protocol's fidelity tool: session completion rates and session frequency. Schools achieving high performance across these implementation indicators will be classified as high-compliance for dose-response analysis. Training completion and coach caseload dimensions operate at programme level and cannot be disaggregated by school.

The evaluation team will receive monitoring data from RFL covering attendance records, session delivery, and programme completion rates at school level. These data will inform school-level compliance classification through established implementation criteria, with schools achieving at least 75% fidelity classified as compliant.

Due to data anonymisation requirements, individual attendance records cannot be linked to survey outcomes. This precludes student-level compliance analysis and Complier Average Causal Effect estimation. Individual engagement patterns will be reported descriptively but cannot be incorporated into causal analyses.

Treatment effects in the presence of non-compliance

We will examine intervention effects among schools that successfully implemented the programme. This involves creating a three-category compliance variable: high-compliance schools (≥75% fidelity), low-compliance schools (<75% fidelity), and control schools (reference category).

Our analytical approach substitutes school-level compliance status for treatment allocation in our fully specified model (Model 3):

$$Y_{ij} = \beta_0 + \beta_1 RFL - Compliance_j + \beta_2 Baseline_{ij} + \beta_3 Wave_j + \beta_4 Strat_j + \mu_j + \varepsilon_{ij}$$

This analysis compares outcomes across the three compliance categories whilst maintaining the same covariate structure and multilevel specification as our primary analysis. If all intervention schools achieve high fidelity, this analysis becomes equivalent to our primary intention-to-treat analysis.

4.10 Intra-cluster correlations (ICCs)

In this trial, schools represent the clustering units. Using an empty multilevel model without covariates, we will compute Intracluster Correlation Coefficients (ICCs) for the baseline measure of the primary outcome (externalising behaviour). The ICC estimation will occur at the school level, which corresponds to our study design's clustering level.

We will fit a two-level random intercept model with children and young people (CYP) at level 1 nested within schools at level 2. The ICC calculation will use the following formula:

$$ICC = \frac{\sigma^2_{School}}{(\sigma^2_{school} + \sigma^2_{CYP})}$$

Where σ^2_{School} represents the variance at the school level, and σ^2_{CYP} represents the variance at the CYP level.

We will utilise the 'estat icc' command in Stata© 17 to derive the ICC estimate and its corresponding confidence interval from the empty multilevel model[8].

Furthermore, we will calculate the ICC from the primary analysis model (which includes covariates and additional predictors) to evaluate how adjusting for these variables affects the clustering effect.

4.11 Presentation of outcomes

As we are using a multilevel model, we will use the effect size for cluster-randomised trials adapted from Hedges (2007) as below:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{(\sigma^2_S + \sigma^2_{error})}}$$

[8] This has already been estimated on baseline data

- $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between both arms adjusted for baseline characteristics;
- $\sqrt{(\sigma_S^2 + \sigma_{error}^2)}$ is the estimated population standard deviation obtained from an 'empty' multilevel model with no predictors.

Therefore, the effect size (ES) quantifies the portion of the population's standard deviation attributable to the intervention (Hutchison & Styles, 2010). Additionally, a 95% confidence interval for all effect sizes, adjusted for the clustering of pupils within schools, will be provided. Effect sizes will be computed for each of the estimated regressions.

5 References

Bosker, R., & Snijders, T. A. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modeling. *Multilevel Analysis*, 1–368.

Caldwell, D. M., Davies, S. R., Thorn, J. C., Palmer, J. C., Caro, P., Hetrick, S. E., Gunnell, D., Anwer, S., López-López, J. A., French, C., Kidger, J., Dawson, S., Churchill, R., Thomas, J., Campbell, R., & Welton, N. J. (2021). School-based interventions to prevent anxiety, depression and conduct disorder in children and young people: A systematic review and network meta-analysis. *Public Health Research*. https://doi.org/10.3310/phr09080

Carril, A. (2017). *RANDTREAT: Stata module to randomly assign treatments uneven treatments and deal with misfits* [Computer software]. https://econpapers.repec.org/software/bocbocode/s458106.htm

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

Donner, A., & Klar, N. (2004). Pitfalls of and Controversies in Cluster Randomization Trials. *American Journal of Public Health*, *94*(3), 416–422.

Gaffney, H., Jolliffe, D., & White, H. (2023). *Knife Crime Education Programmes*.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64–78. https://doi.org/10.1037/1082-989X.2.1.64

Hutchison, D., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. NFER Slough.

Little, R. J. A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, *88*(421), 125–134. https://doi.org/10.1080/01621459.1993.10594302

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Stone, L. L., Janssens, J. M. A. M., Vermulst, A. A., Van Der Maten, M., Engels, R. C. M. E., & Otten, R. (2015). The Strengths and Difficulties Questionnaire: Psychometric properties of the parent and teacher version in children aged 4–7. *BMC Psychology*, *3*(1), 4. https://doi.org/10.1186/s40359-015-0061-8

Turi, E., Tóth, I., & Gervai, J. (2011). [Further examination of the Strengths and Difficulties Questionnaire (SDQ-Magy) in a community sample of young adolescents]. *Psychiatria Hungarica : A Magyar Pszichiatriai Tarsasag Tudomanyos Folyoirata*. https://www.semanticscholar.org/paper/%5BFurther-examination-of-the-Strengths-and-in-a-of-Turi-T%C3%B3th/3d055bcfac47bc195378884afd2f0ebe93f211f1?utm_source=consensus

Wong, K., Morris, S., Wallace, S., Gray, P., & Burchell, E. (2023). *Rugby Football League – Inspiring Futures Educate Mentoring Programme*.

Zhang, J., Quan, H., Ng, J., & Stepanavage, M. E. (1997). Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials*, *18*(3), 204–221.

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413