STATISTICAL ANALYSIS PLAN

Evaluation of the Trauma Informed Schools UK Training and Implementation (TISUK): A cluster randomised controlled trial

Ipsos UK, The University of Kent, TONIC

Principal investigator: Facundo Herrera, Peter Sakis, Dr Jessica Ozan, Dr Amanda Carr, Prof. Simon Coulton

YOUTH ENDOWMENT FUND

Evaluation of the Trauma Informed Schools UK Training and Implementation (TISUK): A cluster randomised controlled trial

Statistical analysis plan

IPSOS UK, TONIC and the University of Kent

Principal investigators: Facundo Herrera, Peter Sakis, Dr Jessica Ozan, Dr Amanda Carr, Prof. Simon Coulton

Project title	*Evaluation of the Trauma Informed Schools UK Training and Implementation (TISUK): A cluster randomised controlled trial*
Developer (Institution)	*TISUK*
Evaluator (Institution)	Ipsos UK, The University of Kent, TONIC
Principal investigator(s)	Facundo Herrera, Peter Sakis, Dr Jessica Ozan, Dr Amanda Carr, Prof. Simon Coulton
SAP author(s)	*Facundo Herrera, Prof. Simon Coulton*
Trial design	Two-armed cluster randomised controlled trial with random allocation at the school level
Trial type	Efficacy
Evaluation setting	School
Target group	Children and young people (CYP) in Year 8
Number of participants	100 schools and 16,000 Year 8 pupils (with 1,100 receiving targeted support)

Primary outcome and data source	Externalising behaviour: SDQ – combined conduct and hyperactivity scores (Survey)
Secondary outcome and data source	*(CYP) Internalising behaviour: SDQ – combined emotional regulation and peer problems (Survey)* *(CYP) Prosocial behaviour: SDQ – Prosocial behaviour (Survey)* *(CYP) Total difficulties: SDQ – combined emotional regulation, hyperactivity, conduct, and peer problems (Survey)* *(CYP) Non-psychotic psychological distress: General Health Questionnaire (GHQ12) (Survey)* *(CYP) Well-being: Short Warwick Edinburgh Mental Well-being Scale (Survey)* *(CYP) Sense of connectedness: School Connectedness Scale (Survey)* *(CYP) Percentage of exclusions & suspensions: administrative records* *(CYP) School attendance: administrative records* *(School staff) Attitudes related to TIC: Attitudes Related to Trauma Informed Care (ARTIC) (Survey)* *(School staff) Well-being: Short Warwick Edinburgh Mental Well-being Scale (Survey)* *(School) Staff retention: administrative records* *(School) Staff sickness: administrative records* *(School) Percentage of exclusions & suspensions of CYP: administrative records* *School) Percentage of school attendance of CYP: administrative records*

SAP version history

Version	Date	Changes made and reason for revision
1.1	July 2024	Updated section on compliance (see section 2.3.9)
1.0 [*original*]	April 2024	*[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

1 Introduction

The intervention is called Trauma Informed Schools UK Training and Implementation. It is a whole-school intervention aimed at improving school staff awareness of trauma-informed practices (TIPs) and implementing changes in policies and practices consistent with TIPs.

The intervention targets school staff, who are expected to implement institutional changes, and all children and young people (CYP) attending the school. The intervention encompasses various activities:

1. Whole staff training: Two 3-hour sessions for all staff, including support staff and administrators, providing an overview of trauma, its impact, adverse childhood experiences (ACEs), protective factors, neuroscience of trauma, and relational approaches.
2. Senior leadership training: A 2-day training session for 5-7 members of the Senior Leadership Team (SLT) from each school, focusing on creating a trauma-informed and mentally healthy culture through ethos, policy, and practice.
3. Network consultancy support: Three consultancy support meetings for school leaders to embed changes in culture, policy, and practice and identify and overcome implementation barriers.
4. Diploma practitioner training: An 11-day Level 5 Diploma Practitioner Training for 5-7 staff members from each school, providing a deeper understanding of trauma and its recovery process and equipping them with skills and knowledge to respond effectively and support the most vulnerable pupils through individual or small group support.
5. Reflective supervision workshops: Training for two practitioners from each school to establish an effective, sustainable reflective supervision model for key staff.
6. Webinar viewing for staff and CYP: Access to webinars on topics related to trauma, relationships, emotions, and gangs/violence, with discussions facilitated by teachers.

The expected outcomes within the first 5-6 months include increased knowledge, understanding, and skills among staff; enhanced staff support structures and changes in school policies and procedures; improved mental health and well-being, reduced exclusions, and increased engagement among students.'

Further information on the intervention can be accessed here: https://www.traumainformedschools.co.uk

This programme will be evaluated through an efficacy trial, a two-arm and a cluster-randomised controlled trial (cRCT). The trial is clustered at the school level as the programme is a whole-school intervention. Every school that signs up to participate in the trial will have an equal probability (50%) of being assigned to the treatment or control groups. In addition, the trial is complemented by an Implementation and Process Evaluation (IPE) and an embedded quasi-experimental design (QED).

The main focus of the process evaluation will be to accurately assess the extent to which the intervention is implemented as intended throughout the school, ensuring fidelity to the intervention's principles. In addition to fidelity, other dimensions to test are dosage, responsiveness, quality, and reach.

The intervention comprises two participation levels for children and young people (CYP). At the whole school level, staff implement structural changes affecting all CYP, with outcomes measured in Year 8. At a targeted level, diploma practitioners provide tailored support to a subgroup of CYP with trauma history. A Study Within a Trial (SWAT) focuses on Year 8 CYP for additional support, employing a quasi-experimental design with propensity score matching (PSM) due to non-randomised control group selection (covered in section 3 below).

2 Main trial on whole school intervention

2.1 Design overview

This efficacy trial adopts a two-arm, two-level design with pupils clustered into schools. The unit of randomisation is the school. All pupils in Year 8 in the participating schools undertake baseline tests from November 2023 to January 2024 and outcome tests at the follow-up stage in March 2025. The schools within the control group continue to function according to business-as-usual. There is no waitlist design in this trial.

The randomisation process adopts the minimisation approach considering whether the percentage of students eligible for free school meals (FSM6) at each school fell above or below the median percentage of FSM6 across all schools in the sample. More detail is discussed in the next section.

The randomisation process was executed by colleagues from the University of Kent, operating independently from the evaluation team at Ipsos UK. The University of Kent was provided with a list of schools without school names but marked with ID codes, ensuring they could not discern the schools' identities, thus maintaining blindness during the randomisation procedure. The randomisation procedure occurred in two phases: one in early December 2023 and another in January 2024. Initially, the intention was to conduct baseline testing for

all schools in November and proceed with randomisation in December. However, extending the timeline became imperative to accommodate additional time for recruitment, data collection, fieldwork, and the constraints posed by the Christmas break and Ofsted inspections, allowing schools ample opportunity to implement the surveys.

The primary outcome is externalising behaviour at pupil level in Year 8 measured by the Strengths and Difficulties Questionnaires (SDQ) – Combined Conduct and Hyperactivity Scale. There is a baseline measurement before the start of the intervention and then a follow-up measurement once the intervention ends, namely, 15 months after randomisation. The secondary outcomes at pupil levels in Year 8 consist of internalising behaviour, prosocial behaviours, total difficulties, non-psychotic psychological distress, well-being, sense of connectedness, exclusions and suspensions, and school attendance. Other secondary outcomes at the school staff level are attitudes related to TIC, school staff's well-being, school staff retention and sickness. Finally, the % of exclusions, suspensions and attendance of CYP at the school level will be collected. Detailed instruments for each outcome are described in Table 1 below.

Table 1 Summary of trial design

Trial design		**Two-arm and cluster randomised at the school level**
Unit of randomisation		**School**
Stratification variables (if applicable)		**Binary indicator of school-level percentage of pupils eligible for Free School Meals (Below median percentage vs. At or above median percentage)**
Primary outcome	variable	**(CYP) Externalising behaviour**
	measure (instrument, scale, source)	**SDQ – Combined conduct and hyperactivity scale scores (0-20) (Survey)**
Secondary outcome(s)	variable(s)	**(CYP) Internalising behaviour** **(CYP) Prosocial behaviour** **(CYP) Total difficulties** **(CYP) Non-psychotic psychological distress**

		(CYP) Well-being **(CYP) Sense of connectedness** **(CYP) Exclusions & suspensions** **(CYP) School attendance** **(School staff) Attitudes related to TIC** **(School staff) Well-being** **(School) Staff retention** **(School) Staff sickness** **(School) Exclusions & suspensions of CYP** **(School) School attendance of CYP**
	measure(s) (instrument, scale, source)	**(CYP) Internalising behaviour:** SDQ – Combination of emotional regulation and peer problems (0-20) (Survey) **(CYP) Prosocial behaviour:** SDQ – sub-dimension of prosocial behaviour (0-10) (Survey) **(CYP) Total difficulties:** SDQ – Combination of sub-dimensions of conduct, hyperactivity, emotional regulation, and peer problems (0-20) (Survey) **(CYP) Non-psychotic psychological distress**: General Health Questionnaire (GHQ12) (0-12) (Survey) **(CYP) Well-being**: Short Warwick Edinburgh Mental Well-being Scale (7-35) (Survey) **(CYP) Sense of connectedness:** School Connectedness Scale (Survey) **(CYP) Exclusions & suspensions:** administrative records **(CYP) School attendance:** administrative records **(School staff) Attitudes related to TIC:** Attitudes Related to Trauma Informed Care (ARTIC) (Survey) **(School staff) Well-being:** Short Warwick Edinburgh Mental Well-being Scale (7-35) (Survey) **(School) staff retention:** administrative records

		(School) staff sickness: administrative records **(School) Percentage of exclusions & suspensions of CYP:** administrative records **(School) Percentage of school attendance of CYP:** administrative records
Baseline for primary outcome	variable	**(CYP) Externalising behaviour**
	measure (instrument, scale, source)	**SDQ – Combined conduct and hyperactivity scale scores (Survey)**
Baseline for secondary outcome	variable	**(CYP) Internalising behaviour** **(CYP) Prosocial behaviour** **(CYP) Total difficulties** **(CYP) Non-psychotic psychological distress** **(CYP) Well-being (CYP) Sense of connectedness** **(CYP) Exclusions & suspensions** **(CYP) School attendance** **(School staff) Attitudes related to TIC** **(School staff) Well-being** **(School) Staff retention** **(School) Staff sickness** **(School) Percentage of exclusions & suspensions of CYP** **(School) Percentage of school attendance of CYP**
	measure (instrument, scale, source)	**(CYP) Internalising behaviour:** SDQ – Combination of emotional regulation and peer problems (0-20) (Survey) **(CYP) Prosocial behaviour:** SDQ – sub-dimension of prosocial behaviour (0-10) (Survey) **(CYP) Total difficulties:** SDQ – Combination of sub-dimensions of conduct, hyperactivity, emotional regulation, and peer problems (0-20) (Survey)

		(CYP) Non-psychotic psychological distress: General Health Questionnaire (GHQ12) (Survey) **(CYP) Well-being**: Short Warwick Edinburgh Mental Well-being Scale (Survey) **(CYP) Sense of connectedness:** School Connectedness Scale (Survey) **(CYP) Exclusions & suspensions:** administrative records **(CYP) School attendance:** administrative records **(School staff) Attitudes related to TIC:** Attitudes Related to Trauma Informed Care (ARTIC) (Survey) **(School staff) Well-being:** Short Warwick Edinburgh Mental Well-being Scale (Survey) **(School) staff retention:** administrative records **(School) staff sickness:** administrative records **(School) Number of exclusions & suspensions of CYP:** administrative records **(School) School attendance of CYP:** administrative records

2.2 Sample size calculations overview

Table 4 describes the minimum detectable effect size (MDES) estimates and sample sizes for the TISUK impact evaluation. The goal was to adopt conservative assumptions to allow for a well powered statistical analysis sufficient to detect an MDES below 0.20 [1] and sub-population analysis.

Sample size estimations and assumptions

The goal is to calculate the required sample size for detecting a meaningful difference in the primary outcome measure of adolescent externalising behaviours. Key parameters involved in the calculation include the effect size, power, alpha level, and accounting for the clustered

[1] All YEF evaluations require a minimum detectable effect (MDES) below 0.20

design. Table 4 describes the sample size calculated using Stata 17© according to the following assumptions:

1. A pre-post zero correlation is assumed for externalising behaviours, as there are no reliable sources from which to base an estimate. Furthermore this represents the most conservative scenario for sample size estimation. Assuming a non-zero correlation would risk underestimating the required sample size if the assumed correlation is too high, we obtain a conservative sample size estimate for a covariance analysis by omitting this parameter and assuming a zero correlation. .

2. The SDQ (externalising scale) adolescent population mean is 6.0, and the assumed pooled Standard Deviation (SD) is 1.74[2], using the UK self-report population norms for 11-15 year olds[3]. Thus, to detect an effect size of 0.2 or greater, we need an SDQ externalising behaviours score difference of 6.0 (control) versus 5.66 (intervention). This difference requires 527 participants in each group, 1,054 for the whole sample.

3. Assuming 10% pupil-level attrition between baseline and follow-up requires a total sample size of 1,172.

Sample Size Calculation

We estimate a required total sample size of 6,798 participants for the proposed cluster randomised controlled trial. This calculation accounts for clustering using the design effect, based on an assumed ICC of 0.03 and a harmonic mean year group size of 200 pupils — assuming an 80% consent rate from the average year group size of 197 in England (ONS, 2023). The design effect is estimated to be 5.8[4], which inflates the initially calculated sample size of 1,172 participants (586 per group for an individually randomised trial) to 6,798. The ICC estimate of 0.03 is based on previous studies by Shackleton, Hale, Bonell, and Viner (2016) and Parker, Nunns, Xiao, Ford, and Ukoumunne (2021), which estimated ICCs for adolescents' health outcomes across secondary school settings in 21 European countries.

[2] The SD has been estimated as the standardised SD using the same raw data for both scales

[3] https://www.sdqinfo.org/norms/UKNorm1.pdf

[4] This is 1 + (0.03 * 160)

Given assumptions 1 to 3 above, we need a minimum of at least 25 schools in each arm for an efficient cluster trial, that is 50 schools in total. Nevertheless, we propose double the sample size aiming to recruit 100 schools with 50 on each arm. The rationale for this is further discussed below.

Table 2 provides an overview of the numbers available at the primary endpoint for a variety of school retention rates (100%, 90%, 80%, 70%, 60%) and pupil retention rates (100%, 90%, 80%, 70% and 60%) scenarios. In all scenarios the numbers recruited and followed up are sufficient to meet the sample size requirements when we account for the different design effects (i.e., as more pupils drop out, the design effect, which is a function of the cluster size, decreases, hence the overall sample size decreases). Table 3 describes the impact on effect size by number of schools and pupils retained.

Table 2 Impact of different pupil and school retention rates on sample size

		Proportion of pupils retained				
		100%	90%	80%	70%	60%
Number of schools retained	100	16,000	14,400	12,800	11,200	9,600
	90	14,400	12,800	11,200	9,600	8,640
	80	12,800	11,200	9,600	8,640	7,680
	70	11,200	9,600	8,640	7,860	6,720
	60	9,600	8,640	7,860	6,720	5,760

Table 3 Impact of different pupil and school retention rates on MDES

		Proportion of pupils retained				
		100%	90%	80%	70%	60%
Number of schools retained	100	0.1064	0.1074	0.1086	0.1102	0.1122
	90	0.1122	0.1132	0.1145	0.1161	0.1183

	80	**0.1190**	**0.1201**	**0.1214**	**0.1232**	**0.1255**
	70	**0.1272**	**0.1284**	**0.1298**	**0.1317**	**0.1341**
	60	**0.1374**	**0.1385**	**0.1402**	**0.1422**	**0.1449**

Rationale for Targeting 100 Schools

Targeting 100 schools and 16,000 students is justified by several considerations. First, a larger pool of schools increases the likelihood of recruiting a diverse population of school types (e.g., school-level of deprivation, FSM, different socioeconomic backgrounds) and student demographics, enabling robust subgroup analyses and generalizability of findings across various contexts. Second, a large overall sample size enhances statistical power to detect smaller but potentially meaningful intervention effects, minimising the risk of false negatives and providing reliable findings to inform future interventions. Third, the target sample size was selected to ensure sufficient power for the quasi-experimental design (QED) analysis focused on a targeted subgroup of students. Lastly, a larger sample size accommodates the cluster randomised trial design, accounting for potential clustering effects and maintaining adequate power for the primary analysis.

Table 4 displays the minimum detectable effect in the protocol according to the assumed parameters, target sample size, and the respective effect size after randomisation driven by the achieved number of responses on the primary outcome. Thus, while the MDES in the protocol is an estimation from the assumptions, the MDES estimated in the column 'Randomisation' corresponds to the actual achieved sample size at baseline in the primary outcome from the schools that were randomised. As we are adopting an intention-to-treat approach, the estimation includes responses from 4 schools that dropped out after randomisation.[5] Those four schools that dropped out belong to the intervention group.

[5] This means that now the number of schools in the trial are 74 with 4 dropping out from the intervention group after randomisation

Table 4 Sample size calculations and effects size

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		**0.1064**	**0.094**
Pre-test/ post-test correlations	level 1 (participant)	n/a	n/a
	level 2 (cluster)	n/a	n/a
Intracluster correlations (ICCs)	level 1 (participant)	n/a	n/a
	level 2 (cluster)	**0.03**	**0.015**
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided
Average cluster size		**160**[6]	**166**[7]
Number of clusters	intervention	50	40
	control	50	38
	Total	100	78[8]

[6] Average school size in England is 986, average year size is 200 and accounting for a potential 20% not consenting, hence, cluster size 160.

[7] This is the actual cluster average size

[8] Although the initial plan was to allocate schools equally to the intervention and control groups at the randomisation stage, the achieved sample showed a slight difference due to attrition during the recruitment

		Protocol	Randomisation
Number of participants	intervention	8,000	6,391
	control	8,000	7,214
	Total	16,000	13,393[9]

The MDES calculations were performed using PowerUp! Tool, developed by Dong and Maynard (2013), which incorporates the formula below according to Bloom, Richburg-Hayes, and Black (2007), for the two-level cluster randomised controlled trials:

$$MDES = M_{n-k*-2}\sqrt{\left(\frac{\rho(1-R_C^2)}{P(1-P)J}\right) + \left(\frac{(1-\rho)(1-R_i^2)}{P(1-P)Jm}\right)}$$

Employing this formula enables the inclusion of baseline covariates at the cluster (school) level, pupil level, or both. Our analysis strategy intends to use mean-cantered baseline scores within the multilevel models at both the pupil and school levels. Incorporating a covariate at both levels - school and pupil - enhances the trial's precision, resulting in a reduced Minimum Detectable Effect Size (MDES) estimate compared to methodologies that employ only one covariate or none.

2.3 Analysis

The analysis was chosen before collecting baseline survey data. The syntax in Stata© is included at the end of this section. The analysis will be conducted using an analysis by intention to treat (ITT) and will include all available data, maintaining participants as members of their allocated group. The following research questions will be addressed:

Pupil level

and baseline testing process. As a result, there is a difference of 2 schools between the intervention and control groups, partly because randomisation occurred over two waves of recruitment.

[9] This is the achieved sample size for the primary outcome (Externalising behaviours-SDQ)

- 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 TISUK training and CYP in control settings receiving business as usual at follow-up?

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

- ERQ3: What is the mean difference in prosocial behaviour, measured by the SDQ subdomain of Prosocial behaviour, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business-as-usual at follow-up?

- ERQ4: What is the mean difference in Total Difficulties, measured by the SDQ subdomain of Conduct Problems, Hyperactivity, Emotional Problems and Peer Problems, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business-as-usual at follow-up?

- ERQ5: What is the mean difference in non-psychotic psychological distress, measured by the General Health Questionnaire (GHQ), between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ6: What is the mean difference in well-being, measured by the Short Warwick Edinburgh Well-being Scale (SWEMWBS), between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ7: What is the mean difference in the sense of connectedness, measured by the School Connectedness Scale, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ8: What is the mean difference in the percentage of exclusions between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ9: What is the mean difference in the percentage of suspensions between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ10: What is the mean difference in the percentage of attendance between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?

- ERQ11: What is the mean difference in all primary and secondary CYP outcomes between CYP in intervention settings who received TISUK training and CYP in control settings who received business as usual, considering sub-group analysis by sex, ethnicity, and free school meal (FSM) eligibility?

School staff level

- ERQ12: What is the difference in attitudes related to trauma-informed care (TIC) of school staff, measured by the Attitudes Related to Trauma-Informed Care (ARTIC) survey, between school staff in the intervention setting receiving TISUK training and school staff in control settings receiving business as usual at follow-up?

- ERQ13: What is the difference in well-being, measured by the SWEMWBS, between school staff in the intervention setting receiving TISUK training and school staff in control settings receiving business as usual at follow-up?

School level

- ERQ14: What is the difference in the percentage of school staff retention at the school level between schools in the intervention setting receiving TISUK training and schools in control settings receiving business as usual at follow-up?

- ERQ15: What is the difference in the percentage of school staff sickness at the school level between schools in the intervention setting receiving TISUK training and schools in control settings receiving business as usual at follow-up?

- ERQ16: What is the difference in the percentage of CYP suspensions at the school level between schools in the intervention setting receiving TISUK training and schools in control settings receiving business as usual at follow-up?

- ERQ17: What is the difference in the percentage of CYP exclusions at the school level between schools in the intervention setting receiving TISUK training and schools in control settings receiving business as usual at follow-up?

- ERQ18: What is the difference in the percentage of CYP school attendance in the targeted years at the school level between schools in the intervention setting receiving TISUK training and schools in control settings receiving business as usual at follow-up?

- ERQ19: What is the difference in the percentage of CYP school suspensions/exclusions/attendance in the targeted years at the school level between schools in the intervention setting receiving TISUK training and schools in control settings

receiving business as usual at follow-up, considering sub-group analysis by sex, ethnicity, and free school meal (FSM) eligibility?

As evidence suggests, the distribution of the externalising behaviour 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 school 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[10] 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).

Secondary outcomes will be assessed similarly by establishing diagnostic plots to explore the distribution and identify the most appropriate regression approach, including stratification factors and baseline covariates within a multilevel model. These plots include residual plots to check for violations like non-linearity and heteroscedasticity, normal probability plots to assess the normality of residuals, and histograms of residuals to identify potential outliers.

Outcomes at the individual level (CYP and school staff) will be analysed using multilevel modelling, while outcomes at the school level will be done through linear regression. Table 5 sets out the statistical analysis for each outcome by level.

Table 5 Statistical analysis by outcome and level

Outcome	Level	Model	Covariates
<u>Primary outcome</u> Externalising behaviour	CYP	Multilevel model	Pre-treatment scores of outcomes at CYP level

[10] This refers to the binary variable indicating whether the school is above or below the median proportion of FSM6 students

Secondary outcomes Internalising behaviour Non-psychotic psychological distress Well-being Sense of connectedness Exclusions, suspensions and attendance at individual (CYP) level			Demographic factors (sex, FSM status) at CYP level
Secondary outcomes Attitudes related to TIC Well-being	School Staff	Multilevel model	Pre-treatment scores of outcomes at staff level
Secondary outcomes Staff retention Staff sickness Exclusions & suspensions of CYP School attendance of CYP	School	Linear regression (OLS)	Pre-treatment scores of outcomes at the school level

The syntax in Stata© would be as below, using the 'mixed' command:

Full adjusted model: mixed post-treatment_outcome TISUK baseline_outcome fsm || school_id

2.3.1 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 employ a two-level multilevel model to account for the clustered nature of our data, where students are nested within schools. This modelling approach assumes that the schools included in the study represent a random sample from the broader population of schools. Multilevel models are well-suited for analysing hierarchical data structures and appropriately handling the variability within schools (among students) and between schools in terms of

outcomes. These models can effectively capture and model the complex sources of variation at multiple levels of the hierarchy (Bosker & Snijders, 2011).

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

$$Y_{ij} = \beta_1 + TISUK_j\tau + \left(Baseline_{ij}\right)\beta_2 + FSM_j\beta_3 + \mu_j + \varepsilon_{ij}$$

- Y_{ij} is the outcome for pupil i in school j;
- $TISUK_j$ is a binary variable denoting whether a school is assigned to the intervention (1) or the control (0);
- $Baseline_{ij}$ represents pupil-level pre-test covariates and demographic characteristics at baseline for pupil i in school j.
- FSM_j is a binary indicator denoting whether a school is above the median proportion of FSM6 (1) or below (0).
- μ_j are the school level residuals [$\mu_j \sim i.i.d\ N(0, \sigma_\mu^2)$]
- ε_{ij} are the individual level residuals [$\varepsilon_{ij} \sim i.i.d\ N(0, \sigma_\varepsilon^2)$]

This is a random intercept model because $\beta_{1j} = \beta_1 + \mu_j$ corresponds to the school-level intercept for school j and $\beta_{1j} \sim i.i.d\ N(\beta_1, \sigma_\mu^2)$. The total variance is decomposed into two portions: the between-school variance (μ_j) and the within-school variance (ε_{ij}). The target parameter of the intervention is the average effect on pupil outcomes captured by τ.

Regarding the risk of having several outcomes, we have designated a primary outcome and several secondary outcomes in this trial, with the primary hypothesis being addressed through the primary outcome measure. While adjustment for multiple testing is a commonly advocated strategy to mitigate the risk of type I error—incorrectly concluding an effect due to a sampling artefact—this approach has the trade-off of potentially increasing the likelihood of type II errors, where a true effect is overlooked. Given the structure of this trial, where a clear distinction is made between the primary and secondary outcomes, strict adjustment for multiple testing may not be necessary. However, it is crucial to discuss the implications of our findings in a nuanced manner in the final report, considering the context of multiple outcomes and the balance between type I and type II error risks. This approach aligns with the understanding that while avoiding false positives is important, it should not come at the cost of missing genuine effects, as discussed in Zhang, Quan, Ng, and Stepanavage (1997).

2.3.2 Secondary outcome analysis

The remaining secondary outcomes at the pupil and school staff levels follow the same equation as above. Secondary outcomes measured at the school level will be estimated using linear regression (ordinary least squares, OLS) with robust standard errors to account for potential heteroscedasticity. Since there is only one observation per school for school-level outcomes, clustering at the school level is not applicable, and robust standard errors are used to ensure valid statistical inferences in the presence of heteroscedasticity.

The **school-level model** is a linear regression set out by:

$$Z_j = \gamma_0 + \gamma_1 Intervention_j + \gamma_2 BaseS_j + \gamma_3 FSM_j + \mu_j$$

$$\mu_j \sim N(0, \sigma_\mu^2)$$

Where:

- Z_j is the outcome at the school level;
- γ_0 is the intercept;
- $Intervention$ is a binary variable that equals 1 if school is within intervention arm and 0 otherwise;
- $BaseS_j$ represents the baseline covariate at the school level, which is centred around the grand mean of the school level;
- FSM_j is a binary variable indicating whether a school's % proportion of FSM pupils is above or below the median in the sample of schools participating in the trial;
- μ_j is the random error across all school
- σ_μ^2 is the residual/error variance between schools

2.3.3 Sub-group analysis

A priori in the protocol has been set out for the sub-group analysis. The sub-group analysis is exploratory and will be twofold. First, we will conduct a latent class analysis (LCA) on the intervention group to explore whether there are emerging sub-groups with differential effects. Second, we will conduct a multilevel model by sub-groups, considering that a model with interaction may be more demanding regarding power analysis. However, we will also consider a multilevel model with interaction between treatment and sub-group dummy variables for robustness check.

Latent class analysis

LCA refers to the model whose underlying indicators are all categorical, while for the continuous case, LCA is known as latent profile analysis (LPA) (Sinha, Calfee, & Delucchi, 2021). The LCA model will include the socio-demographic variables (sex, ethnicity, FSM) and primary and secondary outcomes. This will allow us to observe 'latent' groups emerging from the data.

In latent class analysis (LCA), the probability of an individual belonging to a particular latent class is modelled based on observed categorical indicators. The model assumes that the observed variables are conditionally independent given the latent class. This analysis is described by the below equation, assuming a Y_{ij} observed categorical variable for individual i in group j. The probability of observing a specific pattern of responses Y_{ij} given the latent class c is modelled using a multinomial logistic regression, typically represented as:

$$P(Y_{ij} = k|C_{ij} = c) = \frac{\exp(\lambda_{kj})}{\sum_{l=1}^{K} \exp(\lambda_{lj})}$$

Where:

- $P(Y_{ij} = k|C_{ij} = c)$ is the probability of observing response category k for individual i in group j given latent class c.
- K is the total number of response categories.
- λ_{kj} is the log-odds of individual i in group j selecting response category k when belonging to latent class c.

In the context of this analysis plan:

- The observed categorical variables Y_{ij} will include socio-demographic variables (sex, ethnicity, FSM) and primary and secondary outcomes.
- The latent class variable C_{ij} represents the unobserved class membership.
- will be estimated for each response category and latent class combination.

Multilevel model by sub-groups

The multilevel analysis by groups will explore the differential effects of the intervention by sex, FSM status and ethnicity, described by a modified level-1 equation as below, with φ as the main parameter of interest:

$$Y_{ij} = \beta_1 + TISUK_j\tau + \left(Group_{ij}\right)\beta_2 + \boldsymbol{TISUK_j * Group_{ij}})\boldsymbol{\varphi} + \mu_j + \varepsilon_{ij}$$

2.3.4 Further analysis

Additional analysis will be conducted at baseline after randomisation to check the balance between arms and at the follow-up stage. Incorporating the proportion of pupils with FSM status in the randomisation mechanism is designed to improve the balance across arms. After baseline testing, the evaluation team will check balances across arms. If, by chance, randomisation had not achieved the desired balance on those key socio-demographic characteristics that may affect outcomes, we may consider a multilevel analysis of covariance (ML-ANCOVA) as an additional model to test the robustness of our results. Nevertheless, if the randomisation incorporates % of FSM, it is unlikely that there is an imbalance across arms on FSM.

Stepwise regression analysis will be performed to model the relationship between pre-randomisation factors and demographics on observed outcomes at 15 months. Interaction terms with the allocation arm will be included in the analysis, and a significance level of 0.1 will be used to determine which factors are to be included in the regression model. Pre-randomisation factors will include pupil sex, FSM status and ethnicity. This analysis will be augmented by an additional analysis including participants in the intervention arm using the same pre-randomisation factors, process measures of intervention delivery, and staff changes in perceptions of trauma-informed care (this is for the regression analysis at school staff level).

2.3.5 Interim analyses and stopping rules

There are no interim analysis or stopping rules.

2.3.6 Longitudinal follow-up analyses

There is one follow-up analysis at the end of the intervention for all outcomes, 15 months after randomisation, which follows the outcome analysis described above in sections 2.3.1 and 2.3.2.

2.3.7 Imbalance at baseline

We have employed a minimisation approach to randomisation, which aims to balance the groups by considering the percentage of students eligible for free school meals (%FSM6[11]) at the school level. The minimisation method incorporates %FSM6 as a stratifying factor, ensuring that schools with similar levels of FSM6 are evenly distributed across the control and intervention conditions. While a meticulously executed randomisation process does not guarantee perfect baseline balance, particularly for smaller sample sizes, it increases the likelihood of achieving comparable groups at baseline, as Glennerster and Takavarasha (2013) posited. Given the randomised assignment of schools, any remaining imbalances in baseline characteristics between the groups are expected to be due to chance rather than systematic factors.

Table 6 shows descriptive statistics from the study sample at baseline for the primary outcome and the variable at the school level used for randomisation.

Table 6 Baseline sample balance

School-level (categorical)	**Control group**		**Intervention group**		
	n/N (missing)	**Count (%)**	**n/N (missing)**	**Count (%)**	
FSM6>Median	38	19 (50%)	40	20 (50%)	
Pupil-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size difference**
SDQ - Conduct Problems and Hyperactivity sub-scale	6,179 (2,232)	**7.46 (4.25)**	7,214 (1,983)	**7.62 (4.37)**	0.00

[11] This refers to the situation when a student holds a historical FSM status. Following their FSM eligibility end date, they will retain the "Ever 6 FSM" classification for the subsequent six years. To illustrate, if a student was eligible for FSM from 1st September 2018 until 31st October 2020, their Ever 6 FSM status will continue until 31st October 2026, encompassing a 6-year duration beyond their FSM end date.

2.3.8 Missing data

The proportion of missing data and patterns of missingness will be examined for the primary outcome only, externalising behaviour at six months post-randomisation. Levels of missing data will be reported along with any systematic occurrences of missing data observed in the dataset.

In trials, some participants are inevitably lost to follow-up. The sample size estimation assumed that 20%[12] of participants would not provide an evaluable 6-month follow-up assessment. We will explore the mechanism of missing data to establish whether the data can be considered missing completely at random or missing at random. For each arm, we will present baseline data for those who were followed up at 6 months and those who were lost to follow-up. A logistic regression analysis will be conducted to explore any systematic differences between the allocated groups. If the regression model does not identify any predictor variables, it suggests that the missing data can be considered missing at random.

To avoid loss of efficiency, missing outcome values will be imputed using multiple imputation if the proportion of missing data is greater than 5% and less than 40%. When less than 5% of data is missing, the proportion is considered negligible, and missing observations will be excluded. Multiple imputation methods may perform less well when a substantial amount of data is missing (e.g., more than 40% for the primary analysis). In such cases, the assumptions underlying the imputation become less plausible, and the interpretative limitations of the trial data will be discussed in the results section.

An initial variable reduction analysis will explore the relationship between all potentially prognostic baseline variables and whether a follow-up data point is missing. Only variables where there is an association (p-value > 0.10) will be included in the imputation model.

The number of imputations will be determined to ensure at least 96% statistical efficiency (RE), according to the formula below, where l is the fraction of missing values and M is the number of repetitions:

$$RE = \left(1 + \frac{\lambda}{M}\right)^{-1}$$

[12] This accounts for 10% attrition and 10% non-consent rate

The statistical model and assumptions made in the analysis of the primary outcome will also be implemented in the multiple imputation procedures. If it is suspected that data is missing not at random or if the pattern of missing data is associated with trial allocation, sensitivity analysis will be performed using a pattern mixture approach with mixed modelling and multiple imputation to compare the sensitivity of conclusions to varying assumptions about the missing value mechanism.

2.3.9 Compliance

We will conduct a Complier Average Causal Effects (CACE) analysis using an instrumental variable framework to explore the impact of compliance on the primary outcomes at different levels of compliance. The CACE analysis will be implemented using a two-stage least squares (2SLS) approach clustering the standard errors, where the first stage models compliance as a function of the randomised treatment assignment, and the second stage models the outcome as a function of the predicted compliance from the first stage and other covariates.

Compliance will be measured at the school level through a binary variable, compliance (Yes/No). Compliance for those schools in the intervention arm will be assessed based on the level of engagement. Section 5 discusses the engagement tool in more detail. The tool assesses engagement across five key dimensions:

- Individual Training,
- Whole Staff Training,
- Consultancy,
- Reflective Supervision,
- Webinars.

Each dimension is scored based on specific criteria, such as the number of staff trained, attendance at meetings, and utilisation of resources. The tool uses a point-based system, with 100 points possible across all dimensions. The final engagement score is calculated as a percentage, with clear thresholds for interpretation:

- up to 50% indicates poor engagement,
- 51-75% moderate engagement,
- 76-90% good engagement, and

- 90-100% excellent engagement.

The engagement score will be used to measure compliance in the CACE analysis. CACE weighs the intention-to-treat (ITT) treatment effect by the proportion of compliance, allowing for the estimation of unbiased treatment effects while maintaining the original randomisation in the analysis. Therefore, a school in the intervention arm will be considered compliant with the trial if showing an engagement of 76% or higher. As a robustness check in the analysis, we will explore how compliance is associated with the outcomes and if any threshold for minimum compliance emerges by setting compliance at 76%, 80% and 90% of the engagement tool. This finding could serve in the future of a trial to establish a minimum threshold that ensures a successful delivery.

As set out in the TOC, the impact on CYP is mediated through a change in policies and procedures at the school level. Measuring compliance on that aspect remains challenging. However, this mediator will be assessed qualitatively through the IPE data collection activities.

2.3.10 Intra-cluster correlations (ICCs)

In this trial, the units forming clusters are schools. The Intracluster Correlation Coefficients (ICCs) will be computed for the baseline measure of the primary outcome, externalising behaviour, using an empty multilevel model without covariates. The ICC will be estimated at the school level, as this is the clustering level in the study design.

Specifically, a two-level random intercept model will be fitted, with children and young people (CYP) at level 1 nested within schools at level 2. The ICC will be calculated using the following formula:

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

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

The 'estat icc' command within Stata© 17 will be used to obtain the ICC estimate and its corresponding confidence interval from the empty multilevel model[13].

[13] This has already been estimated on baseline data

Additionally, the ICC will be estimated from the primary analysis model, which includes covariates and other predictors, to assess the impact of adjusting for these variables on the clustering effect.

2.3.11 Presentation of outcomes

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

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

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

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

3 Embedded QED study on targeted intervention

3.1 Design overview

A QED study, using a propensity score matching design, will be embedded within this trial to examine the targeted element of the intervention. In the intervention arm, some staff are given advanced training (Diploma level) to identify a more vulnerable CYP subgroup and provide extra targeted support. This individual intervention is estimated to be provided to about 75 CYP per school, 15 CYP per year group. In the main trial analysis, this group of CYP in Year 8 would become subsumed into the overall intervention group, with all baseline and follow-up data collected on outcomes included in the main trial. Staff who participate in the diploma training will already have identified most young people who would potentially benefit from more intensive support and, in the early stages of training, will be provided with tools to identify any additional young people. By the time of follow-up, at 15 months, this cohort of young people would have been identified and received at least nine months of intervention.

Assuming this group of vulnerable young people is about 15 per school, with a consent rate of 80%, the effective sample size will be 12 per school. Assuming our original aim of 100 schools in the trial, this gives 1,200 CYP across all schools, 600 in the treatment arm matched to 600 in the control arm.

The outcomes for this study will be the same individual pupil-level primary and secondary outcomes as for the main trial, for those that are provided with this individual or small group support in Year 8: the primary outcome is externalising behaviour at pupil level measured by the Strengths and Difficulties Questionnaires (SDQ) – Combined Conduct and Hyperactivity Scale. The secondary outcomes at pupil levels consist of internalising behaviour, prosocial behaviours, total difficulties, non-psychotic psychological distress, well-being, sense of connectedness, exclusions and suspensions, and school attendance.

3.2 Sample size calculations

Sample size estimation and assumptions

The sample size calculation is designed to detect a difference of 0.2, YEF's expected maximum MDES value for efficacy evaluations. To detect this difference with 80% power, an alpha of 0.05 and a two-sided test, at least 527 CYP in each group are required, totalling 1,054 CYP.

For the PSM, the sample size estimations starts from the assumption of 15 CYP on average per school at baseline, assuming a follow-up rate of 80% to account for attrition and consent. The harmonic mean of CYP per year group receiving more intensive support is estimated to be 15. Using this and an ICC of 0.03, the estimated clustered design effect is 1.45. This inflates our required sample to 1,528. With 100 schools, 15 participants per class provide a sample of 1500. Incorporating the inverse propensity score weights will reduce the size of the overall sample by a conservative estimate of 30%, resulting in a required sample of 1,070.

Table 7 indicates the impact of different group sizes (pupils per cluster) on the number of schools needed in the intervention arm of the PSM analysis. The table shows that with 15 pupils per cluster, data from at least 36 intervention schools are required. At the lower end, with 10 pupils per cluster, data from 48 schools would be required. If the follow-up rate is 80% as anticipated, 12 pupils per school will be available, requiring circa 42 schools in each arm of the study. While the target is to maintain 50 schools in each arm, these conservative estimates should be considered.

At the point where the PSM is conducted, the pre- and post-test correlation for the primary outcome can be estimated and incorporated into the sample size calculation, potentially adjusting the required sample size.

Table 7 Impact of differential cluster size on the proportion of schools retained for the intervention group

Pupils	15	14	13	12	11	10
Base sample	527	527	527	527	527	527
Design effect (DE)	1.45	1.42	1.39	1.36	1.33	1.30
Adjusted sample	764	748	733	717	701	685
Sample adjusted for IPW	535	524	513	502	491	480
Number of intervention schools	36	37	39	42	45	48

Table 8 displays the effect size based on the assumption of 12 CYP per cluster at follow-up, assuming 15 at baseline, accounting for attrition and consent. This is a conservative estimate. The second column "Randomisation" displays the effect size assuming 12 CYP per cluster given the number of schools achieved and the observed ICC at baseline. Thus, the total number of participants is an estimation.

Table 8 Sample size calculations - QED

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		0.17	0.20
Pre-test/ post-test correlations	level 1 (participant)	n/a	n/a
	level 2 (cluster)	n/a	n/a
	level 1 (participant)	n/a	n/a

		Protocol	Randomisation
Intracluster correlations (ICCs)	level 2 (cluster)	0.03	0.015
Alpha		0.05	0.05
Power		0.8	0.80
One-sided or two-sided?		Two	Two
Average cluster size (if clustered)		12	12
Number of clusters (schools)	Intervention	50	40
	Control	50	38
	Total	100	78
Number of participants	Intervention	600	456
	Control	600	480
	Total	1,200 (without weighting)	936

3.3 Selection of the comparison group and identification assumptions

An issue that arises is that while we can identify members of this group in the intervention arm of the study by asking for the details of those who receive additional support at six monthly intervals, we cannot identify members in the control arm because there are no set parameters regarding who would be eligible for additional support. As such, we do not have two randomised groups to compare. To address this, we propose to use a matching approach for the quasi-experimental design. Among different matching approaches, Propensity Score Matching (PSM) is the first choice to derive an appropriate group for comparison, although other approaches will be tested for robustness, as further elaborated in section 3.4.3 below.

As Stuart (2010) discussed, exact matching is the ideal scenario, but similar to Mahalanobis distance measures, its performance declines as the number of covariates increases. Requiring exact matches also often results in many unmatched observations, potentially introducing greater bias than allowing more inexact matches. The coarsened exact matching (CEM) approach can perform exact matching on a broader range of variables (Black, Lalkiya, & Lerner, 2020), although its precision decreases with non-informative covariates. Mahalanobis distance matching tends to be effective with a small number of covariates —less than 8, as noted by Rubin (1979) and Zhao (2004)— but its performance suffers when covariates deviate from normality or when dealing with numerous covariates (Gu & Rosenbaum, 1993). This decline likely stems from Mahalanobis matching and treating all interactions among pre-treatment variables with equal importance, which becomes increasingly complex as the number of covariates grows.

Since the literature is inconclusive on which algorithm is superior, given that each has relative strengths and weaknesses, a robust approach is implementing propensity score matching (PSM) along with other matching methods for comparison (Morgan & Winship, 2015). PSM is particularly effective in dealing with many covariates and large samples, increasing the likelihood of finding suitable matches.

The propensity score is the probability of receiving the intervention conditional on measured participant covariates. It is, in essence, a balancing score. If we have two populations, the intervention and control populations, both of which have a similar propensity score, the distribution of baseline covariates will be the same in the intervention and control groups. Hence, we can remove confounding effects by comparing participants who share a propensity score. This is analogous to that induced by randomisation in RCTs.

For the propensity score, the covariates for the matching will be socio-demographic variables such as sex, age, ethnicity, FSM, attendance, exclusions, suspensions and pre-test outcomes for the primary and secondary outcomes. As is often the case in PSM, the selection of the variables will depend on which model delivers the best balance in the matched sample.

A counterfactual control group will be derived using PSM to draw causal inferences of the relative effect of the intervention. A probit regression model will be employed. Callipers of width 0.2 of the standard deviation of the width of the logit propensity score will be employed to maximise matching. Once the propensity scores have been generated, they will be incorporated into the primary and secondary analysis using inverse propensity score weights (IPSW), because this will reduce the sample size required for the control group. One potential limitation of weighting approaches, akin to Horvitz-Thompson's estimation, is the potential for a substantial increase in variance when dealing with extreme weights (i.e. when estimated propensity scores approach 0 or 1). This increased variance is justified if the model is

appropriately specified and the weights are accurate. However, a concern arises when some extreme weights are more closely linked to the estimation process rather than the true underlying probabilities. To mitigate the impact of extreme weights, we would consider weight trimming as a solution, whereby weights exceeding a certain maximum value are capped at that maximum (Stuart, 2010). Robustness check through other matching techniques will allow us to test the sensitivity of this approach to increase in variance. Other matching approaches will include exact matching, Nearest-Neighbour matching, Mahalanobis distance, PSM with different calliper widths, and CEM, as discussed further in section 3.4.3 below.

The success of matching in reducing the imbalance of covariates will be reported by displaying the imbalance before and after matching, that is, between the unmatched and matched samples. The reduction of imbalance will be measured through the reduction in mean and median standardised bias and Rubin's B and Rubin's R. As discussed in Caliendo and Kopeinig (2008), a reduction to a standardised bias between below 2% to 5% is a good sign. Regarding Rubin's metrics, we will consider B < 25 and R between 0.5 and 2 enough balance (Rubin, 2001).

Finally, as acknowledged in the literature, the main limitation of all matching approaches is the inability to control for unobservable covariates (Abadie & Imbens, 2006; Morgan & Winship, 2015). While we have carefully controlled for observable factors through the matching process, there may be unobserved variables, such as student motivation or parental involvement, that could influence the outcome. However, by employing a rigorous matching strategy and drawing from a comprehensive set of observed covariates, we aim to mitigate the potential impact of unobservable confounders as much as possible.

3.4 Analysis

The primary **research question** is:

- 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 TISUK training and CYP in control settings receiving business as usual at follow-up?

The secondary **research questions** are:

- What is the mean difference in internalising behaviour, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomains of Emotional Problems and Peer Problems, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business at follow-up?

- What is the mean difference in prosocial behaviour, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomain of Prosocial behaviour, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business-as-usual at follow-up?
- What is the mean difference in Total Difficulties, measured by the Strengths and Difficulties Questionnaire (SDQ) subdomain of Conduct Problems, Hyperactivity, Emotional Problems and Peer Problems, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business-as-usual at follow-up?
- What is the mean difference in non-psychotic psychological distress, measured by the General Health Questionnaire (GHQ), between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in well-being, measured by the Short Warwick Edinburgh Well-being Scale (SWEMWBS), between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in the sense of connectedness, measured by the School Connectedness Scale, between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in the percentage of exclusion between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in the percentage of suspensions between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in the percentage of attendance between CYP in intervention settings receiving TISUK training and CYP in control settings receiving business as usual at follow-up?
- What is the mean difference in all primary and secondary CYP outcomes between CYP in intervention settings who received TISUK training and CYP in control settings who received business as usual, considering sub-group analysis by sex, ethnicity, and free school meal (FSM) eligibility?

The analysis will be conducted using an intention to treat (ITT) and include all available data, maintaining participants as members of their allocated group.

3.4.1 Primary analysis

The primary analysis will likely be a linear regression model adjusted for baseline stratification covariates (proportion FSM), and the baseline value of the outcome. As there is variation in

business as usual (BAU) across sites, a multilevel model will be estimated to account for pupils being nested within schools. Individual outcomes will be incorporated into the model with an inverse propensity weight for each participant.

3.4.2 Inference

Uncertainty will be reported using confidence intervals, standard errors of estimates and the corresponding p-values.

3.4.3 Robustness Checks

Several robustness checks will be implemented, including the following:

- We will test the common support in the matched sample and the t-test after matching to analyse if the matching is satisfactory, including reviewing the B and R – Rubin to ensure that the matching matches at least well.
- Use Rosenbaum (2002) to evaluate to what extent the ATE estimate is robust to non-compliance with the assumption of selection in observables.
- Measure the sensitivity to the matching method to test the robustness of the main matching technique one-to-one. Different forms of matching will be estimated: (a) calliper (with different widths), (b) nearest neighbour (NN) (tested with five neighbours), and finally, (iii) kernel. The choice of the matching method will be transparently discussed and linked to the present evaluation.
- Estimate the results using the Inverse Probability Weighting (IPW) methodology and Entropy Balancing.
- We will first apply the balance test proposed by Smith and Todd (2005) based on the regression of the covariates against the treatment variable (T) and its polynomial forms, where p(X) rejecting the null hypothesis would imply that the variables Xj are unbalanced between the groups. Second, we will implement a Kolmogorov-Smirnov test to ensure a balance between both groups.
- We will consider, if applicable, the falsification test proposed by Lee and Lemieux (2010)). This test seeks to test the hypothesis that the average effect is zero in covariates or in pseudo-outcomes on which - by definition - no effect of the treatment.

3.4.4 Secondary analyses

Secondary outcomes will be assessed similarly by examining the variable distribution by establishing diagnostic plots to identify the most appropriate regression approach, including

stratification factors and baseline covariates within a multilevel model. Diagnostic plots are graphical tools used to assess the assumptions and fit of a statistical model. These may include residual plots (e.g., residuals vs. fitted values, normal Q-Q plots) to check for non-linear patterns, unequal variance, outliers, and normality assumptions. Partial regression plots can be used to examine the relationship between the outcome and specific predictors, while leverage plots help identify influential observations. Additionally, added variable plots can assess the linearity assumption between the outcome and predictors.

Based on the patterns observed in these diagnostic plots, we may consider transformations, robust regression techniques, or alternative model specifications to address potential violations of assumptions. The stratification factors, such as the binary indicator for the percentage of pupils eligible for free school meals, and baseline covariates, like the baseline measure of the outcome variable and other relevant demographic or school-level characteristics, will be included as predictors in the multilevel model to account for clustering and adjust for potential confounding effects.

3.4.5 Subgroup analyses

Sub-group analysis will be conducted to estimate how the treatment effects vary within groups. This means estimating heterogeneous effects, namely, conditional average treatment effects (CATE). The groups consist of sex, ethnicity and FSM. We will aim to minimise the number of strata within each category whenever applicable and necessary. For example, ethnicity will be divided into minority and non-minority to maximise statistical power.

After passing the balancing test and robustness checks, the approach will first achieve a matched sample, thus obtaining the average treatment effect on the treated (ATET). Then, we will condition the ATET on the respective group variables to obtain treatment effect by stratum, namely the CATE.

3.4.6 Further analyses

There is no further analysis beyond the sub-group analysis discussed above and the robustness checks.

3.4.7 Treatment effects in the presence of non-compliance

We will not conduct any Complier Average Causal Effects (CACE) analysis for this strand of the study. The nature of the targeted support for this sub-group presents significant challenges in measuring compliance, such as the whole school intervention. This targeted intervention is highly personalised and adaptable to the individual needs of each student, which means that the intensity and frequency of engagement can vary greatly.

3.4.8 Missing data

As the generation of intervention and comparison groups are based on complete data, we do not account for missing data.

3.4.9 Presentation of outcomes

This is the same as for the main analysis for the efficacy study—see section 2.3.11 above.

4 References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica, 74*(1), 235-267.

Black, B. S., Lalkiya, P., & Lerner, J. Y. (2020). The trouble with coarsened exact matching. *Northwestern Law & Econ Research Paper Forthcoming*.

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

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., . . . French, C. (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*.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys, 22*(1), 31-72.

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.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420.

Hedges, L. V. (2007). Effect sizes in cluster-randomised designs.

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

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature, 48*(2), 281-355.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*: Cambridge University Press.

ONS. (2023). Schools, pupils and their characteristics - Academic year 2022/23. Retrieved from https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics

Parker, K., Nunns, M., Xiao, Z., Ford, T., & Ukoumunne, O. C. (2021). Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review. *BMC Medical Research Methodology, 21*(1), 1-17.

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

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*(366a), 318-328.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188.

Shackleton, N., Hale, D., Bonell, C., & Viner, R. M. (2016). Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM-population health, 2*, 217-225.

Sinha, P., Calfee, C. S., & Delucchi, K. L. (2021). Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. *Critical care medicine, 49*(1), e63.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics, 125*(1-2), 305-353.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics, 25*(1), 1.

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.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of economics and statistics, 86*(1), 91-107.

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413