STATISTICAL ANALYSIS PLAN

Police in Classrooms Randomised Trial (Efficacy)

King's College London and Cardiff University

Principal investigator: Michael Sanders

YOUTH ENDOWMENT FUND

Police in Classrooms Randomised Controlled Trial (Efficacy)

Statistical analysis plan

Evaluating institution: The Policy Institute at King's College London and Children's Social Care Research and Development Centre (CASCADE) at Cardiff University

Principal investigator: Michael Sanders

YEF statistical analysis plan

Project title	Police in Classrooms Randomised Controlled Trial (Efficacy)
Developer (Institution)	PSHE Association
Evaluator (Institution)	King's College London and Cardiff University
Principal investigator(s)	Michael Sanders
SAP author(s)	Michael Sanders, Julia Ellingwood, and Kira Ewanich
Trial design	Two-armed cluster randomised trial with randomisation at the school-year level, stratified by school. The final dataset will include combined data from both the internal pilot and the main efficacy study.
Trial type	Efficacy
Evaluation setting	Secondary schools
Target group	Pupils enrolled in mainstream secondary schools, years 7-10
Number of participants	23,062 (16,262 pupils from the efficacy trial + 6800 pupils from the pilot trial) across 29 schools (20 from the efficacy trial + 9 from the pilot trial)

Primary outcome and data source	Emotional and behavioural difficulties (Strength and Difficulties Questionnaire total difficulties score, captured in pupil self-report)
Secondary outcome and data source	• Offending behaviour and victimhood (Administrative data from constabularies reporting instances of offending and victimhood among pupils enrolled in trial schools) • Delinquent beliefs (Delinquent Beliefs Scale (Thornberry, 1994), captured in pupils survey) • Trust and confidence in police - attitudes, perception of bias, and combined (Perceptions of Police Scale (POPS) (Nadal & Davidoff, 2015), pupils survey) • Disclosure and help-seeking behaviour (bespoke pupil self-report questions) • Deterrence (change in behaviour) (bespoke pupil self-report questions) • School attendance (school administrative data)

SAP version history

Version	Date	Changes made and reason for revision
1.3 [latest]	June 26, 2025	Evaluator responding to peer review comments
1.2	May 9, 2025	Evaluator responding to YEF comments
1.1		
1.0 [*original*]	Update	

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

Table of contents

Introduction

The police in classrooms (PiCl) intervention is the formal delivery of a newly developed Personal, Social, Health, and Economic (PSHE)-written curriculum, taught in classrooms by trained police officers in partnership with teachers. The PSHE curriculum comprises four taught units—Personal Safety, Drugs and the Law, Violence Prevention, and Knife Crime—with each unit containing three lessons. As per PSHE guidance, each unit will be taught collaboratively, with the classroom teacher teaching the first and third lessons within the unit and the specially trained schools officer teaching the middle lesson. The intervention is delivered to pupils in PSHE lessons across four year groups; Y7 – Y10 (age 11-15).

In order to better understand the impact of PiCl on pupil behavioural difficulties (whether it reduces, increases, or leads to no reduction thereof), we are randomising the treatment allocation of a PSHE curriculum at the year-group level (e.g. within a particular school, years 7 and 10 receive treatment, and years 8 and 9 do not, etc). Therefore, within schools involved in the trial, all will have some PiCl implementation but not every year will be within the treated group, depending on how the year group is randomly allocated. For year groups that are allocated to treatment, we will aim for all classes within that year group to receive the PSHE curriculum. This year-group-level randomisation has several benefits. Most notably, it allows us to detect the impact of the intervention with fewer schools than would be required under a school-level randomisation design. It also helps control for school-level confounders and variation in school culture or policies.

The study is designed as a clustered randomised controlled trial, with year groups allocated to treatment or control conditions in a way that ensures balance across the overall sample. Each school will be able to start at a different date as needed, and will run the intervention over the course of an academic term (of which there are three in each school or calendar year for our purposes).

The purpose of the quantitative analysis is to answer the following research questions:

RQ1: What is the impact of receiving the police in classrooms curriculum on emotional and behavioural difficulties, as measured by the self-report Strengths and Difficulties Questionnaire total difficulties score, among secondary school pupils compared with similar pupils who have not received the police in classrooms curriculum?

RQ2: What is the impact of the police in classrooms curriculum on pupils' rates of offending and victimhood, as measured through police administrative data?

RQ3: What is the impact of the police in classrooms curriculum on pupils' beliefs regarding risky or illegal behaviour, as measured through the Delinquent Beliefs Scale?

RQ4: What is the impact of the police in classrooms curriculum on pupils' attitudes to and trust towards the police, as measured through the Perceptions of Police Scale (POPS)?

RQ5: What is the impact of the police in classrooms curriculum on pupils' self-reported confidence in seeking help from police officers?

RQ6: What is the impact of the police in classrooms curriculum on pupils' self-reported behaviours relating to deterrence?

RQ7: What is the impact of the police in classrooms curriculum on pupil attendance among the treated year groups, compared with control year groups and prior years?

RQ8: For the outcomes listed in RQs 1-7, is there heterogeneity in the effectiveness of the intervention for pupils identifying as Black[1], compared with pupils identifying as White?

Design overview

Trial design, including number of arms		Two armed, cluster randomised efficacy trial, with 1:1 treatment allocation ratio
Unit of randomisation		School Year
Stratification variables (if applicable)		School
Primary outcome	variable	Emotional and behavioural difficulties, as measured by the total difficulties score
	measure (instrument, scale, source)	Instrument: Strengths and Difficulties Questionnaire (SDQ) (Goodman et al, 1998) Scale: Total difficulties score, ranging from 0-40 (sum of the SDQ scores from the emotional,

[1] "Black" refers to a range of ethnic sub-groups, including Black African, Black British, Black Caribbean, and any other Black background.

		conduct, hyperactivity, and peer problems subscales) Source: Pupil self-report
Secondary outcome(s)	variable(s)	1. Offending behaviour and victimhood 2. Delinquent beliefs 3. Trust and confidence in police (attitudes, perception of bias, and combined) 4. Disclosure and help-seeking behaviour 5. Deterrence (change in behaviour) 6. School attendance
	measure(s) (instrument, scale, source)	1. Offending behaviour and victimhood (Administrative data from constabularies reporting instances of offending and victimhood among pupils enrolled in trial schools) 2.Delinquent beliefs (Delinquent Beliefs Scale (Thornberry, 1994), captured in pupil self-report) 3. Trust and confidence in police - attitudes, perception of bias, and combined (Perceptions of Police Scale (POPS) (Nadal & Davidoff, 2015), pupils survey) 4.Disclosure and help-seeking behaviour (bespoke pupil self-report questions) 5.Deterrence (change in behaviour) (bespoke pupil self-report questions) 6. School attendance (school administrative data)
Baseline for primary outcome	**variable**	Baseline emotional and behavioural difficulties, as measured by the total difficulties score
	measure (instrument, scale, source)	Instrument: Strengths and Difficulties Questionnaire (SDQ) (Goodman et al, 1998)

		Scale: Total difficulties score, ranging from 0-40 (sum of the scores from the emotional, conduct, hyperactivity, and peer problems subscales on the SDQ) Source: Pupils survey
Baseline for secondary outcome	**variable**	Pre-trial levels of offending at school/year group level and baseline pupil self-report measures
	measure (instrument, scale, source)	Police administrative data Pupil self-report

Sample size calculations overview

In the table below, we provide the scenario where we are powering to detect subgroup effects among Black pupils within the year group, utilising a relatively high assumed ICC among Black pupils (rather than the year group total enrolments) and assuming we are recruiting schools with an 8% average enrolment of Black pupils.[2]

The actual sample size and MDES at randomisation will be provided once the final number of schools in the trial is confirmed. We plan to amend the Statistical Analysis Plan (SAP) in September 2025, in the columns labelled "Post-randomisation" in the table below, after we have a clear understanding of how many schools we have managed to recruit to the trial. This update will reflect the actual sample size and MDES, using the same assumptions outlined during the protocol stage

[2] This is based on the estimated percent enrolments of Black pupils among the schools already recruited for the study as of 25 June 2025.

		Protocol – All pupils	Protocol – Black pupils	Post-randomisation – All pupils	Post-randomisation – Black pupils
Minimum Detectable Effect Size (MDES)		0.2	0.2		
Pre-test/ post-test correlations	level 1 (participant)	0.732	0.733	0.732	0.733
	level 2 (cluster)	N/A			
Intracluster correlations (ICCs)	level 1 (participant)	N/A			
	level 3 (cluster)	0.25			
Alpha		0.05			
Power		0.8			
One-sided or two-sided?		Two-sided			
Average cluster size (year group)		200	16		
Number of clusters	intervention	58	58		
	control	58	58		
	total	115	115		
	Intervention	11,531	928		

		Protocol – All pupils	Protocol – Black pupils	Post-randomisation – All pupils	Post-randomisation – Black pupils
Number of pupil participants (assuming average 8% Black pupil enrolments)	Control	11,531	928		
	Total	23,062	1844.8		
Number of required/recruited schools - Total		28.8			
Number of pilot schools		9		9	
Number of required/recruited schools - Efficacy		20			

The sample size for this study was determined a priori and informed by the pilot trial findings. Within the protocol, four different scenarios were considered, taking into account both the overall sample, differing attrition rates, differing ICCs, and subgroup analyses for Black pupils (Sanders et al., 2024a). The table above presents the sample size for scenario 4 in the protocol, which was determined based on the following key assumptions:

- An MDES of Cohen's D = 0.2 was assumed based on the average effect sizes of successful interventions (i.e. interventions that had an effect size greater than zero or close to zero) carried out in schools funded by the EEF, which aimed to narrow the educational attainment gap between the most and least advantaged children in British schools (Sanders et al, 2020).[3]

[3] While a Cohen's D of 0.2 is imperfect and perhaps too high, with some studies putting 0.1 SD as a medium effect size for educational trials as in Kraft, 2019 and Sanders et al, 2020, we are aiming to balance the risk of underpowering against the possibly excessive cost of running a trial powered to detect very small effect sizes.

- An assumed ICC of 0.25 was used to account for the possibility that outcomes are more strongly correlated within the subgroup of Black pupils compared with the entire year group.[4]
- A potential participant attrition of 15% was estimated based on the pilot trial.
- To detect subgroup treatment effects for Black pupils, additional sample size considerations were made and Black pupils were treated as a distinct cluster within year groups, rather than considering total year group enrolments.
- Based on national data, the study assumes an average school year size of 200 pupils, drawn from a birth cohort of approximately 700,000 and 3,500 secondary schools. Pilot data from 9 Bristol-area schools showed an average year group size of 193.8 pupils (median = 177.2).
- An individual-level correlation of 0.733 between pre- and post-test measures was taken from the pilot data collection (this was calculated just for Black pupils, though the estimation for the general sample was very similar at 0.732). This accounts for participant pre/post-test reliability. The final ICCs and pre/post-test correlations will be included in the efficacy report.

All power calculations were conducted in R using the *pwr* package. The primary population of interest is pupils enrolled in mainstream secondary schools, years 7-10.

Analysis

All analyses for the primary and secondary outcomes will be conducted using an intention-to-treat (ITT) approach.

Primary outcome analysis

The methods of analysis have been chosen a priori. The analysis will be run in R 4.5.0 and Stata 18. Statistical significance will be based on the construction and interpretation of 95%

[4] 0.25 is the assumed ICC for Black pupils within each year group (ie the cluster), describing the degree to which they covary in their outcomes compared with their non-Black peers. Evidence suggests that for social-emotional outcomes, classroom- and year-group ICCs are relatively low across all pupils, ranging from 0.0 to 0.05, but with outliers up to 0.2 (Parker et al, 2025). Some limited evidence concludes that certain health-related behaviours correlate more strongly among Black pupils, with an estimated 0.12 (Siddiqui et al, 1996). We have intentionally chosen a relatively high assumed ICC because we suspect that during these years, pupils will tend to self-select into peer groups based in part on ethnicity (Kogachi & Graham, 2021), and that within these peer groups, SDQ scores will exhibit less variation. Note that we did calculate ICCs for the full sample and Black pupil sample from the pilot, but did not find any indication of cluster correlation (approximately 0.01 correlation for the full sample, insignificant); that said, we would like to remain conservative here, especially given the limited data we have from the pilot.

confidence intervals around estimated effects. While p-values will be reported as continuous probabilities, no strict threshold (e.g., p-values < 0.05) will be used to determine significance. Instead, results will be interpreted in the in the context of the magnitude, direction, and precision of estimated effects, as well as their policy and practical relevance. The primary outcome is emotional and behavioural difficulties, which will be measured using total difficulties score from the SDQ in the pupil baseline and endline surveys.

We will analyse data from the pilot trial alongside data collected from the efficacy trial, combining the datasets into a pooled sample. The total dataset will consist of pupils who participated in both the YEF pilot trial and the efficacy trial (Sanders et al., 2024a; Sanders et al., 2024b). This is possible because we will be repeating the use of the SDQ questions, and thus we will have a 1:1 match on these outcomes between the pilot and efficacy trial datasets. datasets.

We estimate an ANCOVA model equivalent to an AR(1) structure with two time points, using ordinary least squares (OLS) regression. The unit of intervention assignment is the year level, while the unit of analysis is the individual student. To address non-independence between individuals, the model will cluster standard errors at the school-year-time-period level using cluster-robust standard errors with the HC2 correction.

The model is be described as follows: We plan a regression model being estimated of the form;

$$O_{iyst} = \alpha + \beta_1 W_{yst} + \beta_2 O_{iyst-1} + \Gamma X_i + \beta_3 M_i + \beta_4 P_i + \delta_s + u_{yst}$$

Where

O_{iyst} is the value of the outcome measure (pupils' self-reported total difficulties score) for i in year y in school s at time t t.

α is a regression constant
W_{yst} is the teatment indicator (1 = intervention, 0 = control)

O_{iyst-1} is the lagged value of the outcome measure for participant i from year y in school s. This value is set to 0 where missing.

X_i is a vector of participant demographic characteristics collected from the pupil self-reports. These include age, sex, ethnicity, and disability.[5] The table below outlines how ethnicity will be constructed.

M_i is a binary indicator of the missingness of participant i's baseline data, set to 1 if missing and 0 else

P_i is a binary indicator for the pilot vs. efficacy phase (1 = pilot, 0 = efficacy)

δ_s is a fixed effect indicator variable for each school s

u_{yst} is an error term clustered at the level of the yeargroup within school and time period[6]

Operationalising Ethnicity

During the pilot stage, we explored using collapsed ethnicity categories in order to operationalise an ethnicity subgroup analysis. We acknowledged that collapsing categories must be done carefully, to avoid flattening differences in experience and perspective between minority ethnic groups, as could be the case if we were to use a blunt Black Asian Minority Ethnic (BAME) category (Aspinall, 2020; Selvarajah et al. 2020). For further justification of the approach to subgroup analysis by ethnicity, see page 14 of the evaluation protocol.

Police IC Codes	Census 2021 (8a)	Survey response	Subgroup Category
IC1/IC2 - White (North and South European)	4. White: English, Welsh, Scottish, Northern Irish or British 5. White: Irish	White British White Irish Any other white background	White

[5] We have chosen to include disability as a key demographic characteristic. During the pilot, a fair number of pupils reported autism and ADHD diagnoses, which both have relationships with behaviour and contact with the criminal justice system (Young et al, 2011; Collins et al, 2023).

[6] Observations within a school year lack independence from eachother, so to address this, we are clustering our standard errors at the level of the cluster and time period (Abadie et al, 2022). We could run a random effects model, but this is likely attenuate effect sizes, which is risky given that we anticipate relatively small effects.

IC3 – Black	2. Black, Black British, Black Welsh, Caribbean or African	Caribbean African Any other Black background White and Black Caribbean White and Black African	Black
IC4 – Indian subcontinent	1. Asian, Asian British or Asian Welsh	Indian Pakistani Bangladeshi	South Asian
IC5 – Chinese/Japanese/Korean/or other Southeast Asian	N/A	Chinese Any other Asian background	East Asian
IC6 – Arab or North African	N/A	Arab	Arab
IC9 – Unknown	7. Other ethnic group	Any other/Not Stated/Prefer not to say	Not stated
N/A	3. Mixed or Multiple ethnic groups 6. White: Gypsy or Irish Traveller, Roma or Other White	White and Asian Any other mixed Gypsy or Irish Traveller	Other

	7. Other ethnic group		

Secondary outcome analysis

As with the primary outcome analysis outlined above, the methods of analysis for the secondary outcomes have been chosen a priori. The analysis will be conducted in R 4.5.0 and Stata 18. Statistical significance will be based on the construction and interpretation of 95% confidence intervals around estimated effects. While p-values will be reported as continuous probabilities, no strict threshold (e.g., p-values < 0.05) will be used to determine significance. Instead, results will be interpreted in the context of the magnitude, direction, and precision of estimated effects, as well as their policy and practical relevance. The table below outlines the secondary outcomes, their measurement instruments, and the source of the data.

Secondary Outcome	Measurement Instrument	Data Source
Offending behaviour and victimhood	-Offense binary, including: violence against the person, possession of weapons, robbery, drug offences, sexual offences -Victim binary	Police administrative data
School attendance	School attendance rates by year group, sex, and ethnicity	School administrative data
Delinquent beliefs	Delinquent Beliefs Scale (Thornberry, 1994)	Pupil self-report
Pupils' trust and confidence in police (attitudes, perception of bias, and combined)	Perceptions of Police Scale (POPS) (Nadal & Davidoff, 2015)	Pupil self-report
Pupils' confidence in seeking help from police	Bespoke pupil self-report questions	Pupil self-report

Deterrence (change in behaviour)	Bespoke pupil self-report questions	Pupil self-report

We will combine data from the pilot trial with data from the efficacy trial into a pooled sample for analysing the trust and confidence of police outcome and the offending and victimhood outcome. The total dataset will consist of pupils who participated in both the YEF pilot trial and the efficacy trial (Sanders et al., 2024a; Sanders et al., 2024b). This is possible because we will be repeating the use of the POPS scale, and thus we will have a 1:1 match on these outcome measures between the two datasets.

This should also be possible for the police data, but with caveats which are at present unknown. The police forces we are working with on the efficacy trial have confirmed that they collect crime data in roughly a similar fashion and that we should be able to map crime categories between Avon and Somerset's recording practices and those of another police force.

We are limited in our ability to incorporate pilot data on help-seeking and deterrence outcomes, as those questions were only included for treated year groups in the endline survey. Therefore pilot data on these outcome measures will not be included in the regression modelling, although descriptive summaries will still be possible.

Offending behaviour and victimhood

To analyse this outcome using police administrative data, we will create a constructed dataset that combines school enrolment totals from the school—providing the number of pupils in each year group, categorized by gender and ethnicity—with detailed records of each incident of offending behaviour or victimhood reported within the police administrative data. Each row in the constructed dataset will represent all individual pupils enrolled in the sampled schools, complete with gender and ethnicity, age, school enrolment, a dummy indicator of whether that individual was involved in a crime as an offender or victim, and then if so, a short record of the crime, including the date and category of crime.[7] We plan to collect this

[7] For most of these rows, the crime dummy variable will be zero, since the vast majority of pupils will not have been involved in a crime. Importantly, these records will be fully anonymous, and we are suppressing cell sizes less than 3, and never at any point do we as researchers get access to the names of the pupils who are involved in crimes.

for one year preceding the intervention and the year following the intervention, allowing us to control for school-school-year level historical data.

Analysis of the police administrative data will be conducted using logistic regression analysis with separate modelling for each binary outcome (offence = [1, 0], victim = [1,0]), using separate datasets (one for suspects, one for victims) derived from school-age-group level.

Significance will be determined based on the construction and interpretation of 95% confidence intervals around estimated effects. While p-values will be reported as continuous probabilities, no strict threshold (e.g., p-values < 0.05) will be used to determine significance.Instead, results will be interpreted in the context of the magnitude, direction, and precision of estimated effects, as well as their policy and practical relevance. As a secondary outcome measure, we have not powered the study to detect an effect at 0.8 probability; due to the nature of the measure being a relatively rare event, it is likely that we will be underpowered in the police data analysis.

We will use the following regression model;

$$\text{logit}(P(O_{iyst})) = \alpha + \beta_1 W_{yst} + \beta_2 S_s + \beta_3 Y_y + \beta 4 T_t + u_{yst}$$

Where

O_{ist} is the value of the outcome measure for pseudo-individual i in year y in school s at time t

α is a regression constant

W_{st} is a binary indicator of whether or not the year y in school s is treated in time t

S_s is a vector of school level fixed effects

Y_y is a vector of school year fixed effects.

T_t is a binary indicator of time set to 1 in the trial period and 0 else.

u_{ys} is an error term clustered at the level of the yeargroup within school and time period.

School attendance

School attendance rates by year group, sex, and ethnicity will be collected directly from schools. Since school attendance data will be structured very similarly to the offending and victimhood outcome, wherein we will have rates of truancy of pupils by year group, sex,

ethnicity, and school year, we will structure the analysis in the same way as outlined above in the 'Offending and Victimhood' section, using a constructed dataset that combines enrolment numbers and truancy, and modelling individuals' likelihood to be truant.

Delinquent beliefs

The Delinquent Beliefs Scale instrument asks how wrong it is to do 8 different unlawful activities (e.g. "Steal something worth £100"). This measure generates a score up to 32, with higher scores corresponding to stronger beliefs that the activities are wrong. This measure will be analysed as a continuous outcome measure using the same individual level autoregressive (AR(1)) model as the primary outcome.

Pupils' trust and confidence in police (attitudes, perception of bias, and combined)

This measure will be collected from the pupil baseline and endline surveys. The pupil self-report will operationalize trust and confidence in police using the using age-adapted questions from the Perceptions of Police Scale (POPS) (Nadal & Davidoff, 2015). The POPS questionnaire is made up of 12 Likert-scale questions, with lower scores indicating positive views of police and higher scores indicating more negative. The POPS can further be divided into two subscales, with 9 questions corresponding to attitudes about the police (e.g. "The police provide safety") and 3 questions about perceptions of police bias (e.g. "Police officers treat all people fairly"), which will be reported as mean scores. As with the primary outcome, this measure will be estimated using an ANCOVA model equivalent to an AR(1) structure with two time points, using ordinary least squares (OLS) regression.

Pupils' confidence in seeking help from police

This measure will be collected using two bespoke questions in the pupil self-reports:

1. *"If you were worried about something, how likely are you to approach a police officer for help?"*
2. *"How much do you trust the police to handle students' concerns fairly?"*

Responses will be recorded on a Likert scale. This measure will be analysed as a continuous variables using the same methodological approach outlined in the primary outcome section. If the Likert scale data does not meet the model assumptions, which can be determined by checking the data distribution, normality, and number of observations in each category, we will provide descriptive statistics for this outcome.

Deterrence (change in behaviour)

This measure will be collected using two custom statements in the pupil self-reports:

1. *"Having a police officer in school discourages students from misbehaving or breaking school rules."*
2. *"Having a police officer in school discourages students from more serious offending or breaking the law."*

Responses will be recorded on a five-point Likert scale from 'Strongly Disagree' to 'Strongly Agree'. This measure will be analysed as a continuous variables using the same methodological approach outlined in the primary outcome section. If the Likert scale data does not meet the model assumptions, which can be determined by checking the data distribution, normality, and number of observations in each category, we will provide descriptive statistics.

Subgroup analyses

As specified a priori in the protocol, we will conduct exploratory analysis of heterogenous treatment effects among pupils identifying as from a Black ethnic background, compared with pupils identifying as non-Black. . These subgroups will be analysed separately through interaction terms — year group x treatment, and Black x treatment — for the following outcomes: SDQ scores, offending behaviour and victimhood, delinquent beliefs, trust and confidence in the police, and confidence in seeking help from police.

However, we need to manage trade-offs between precision of racial/ethnic subgroups and maintaining statistical power of the subgroup analysis; if we run an interaction across all 18 ethnicities, these small groupings will reduce our overall power and make interpretation opaque. Given this, alongside evidence gathered from the academic literature and recent reports highlighting broader racial disparities in how Black and mixed ethnicity children and young people interact with the criminal justice system—including, but not limited to disproportionate use of strip searching n (Patel, 2020; Runnymede Trust, 2024; YEF, 2025; Yesufu, 2013)— we have opted to use a binary ethnicity measure of Black and non-Black in our subgroup analysis. By being selective in choosing the ethnicity which we are most concerned about in terms of police harm, we improve our statistical power and potential interpretability of our interaction models.

This does have the effect of flattening diverse experiences in the subgrouping of pupils not identifying as Black, however. We think it is likely that other racial and ethnic minority pupils will have differing treatment effects from White pupils and putting them into the same category is simplifying things to a fault. Therefore, we are also planning to explore interaction effects for other ethnicities as well, e.g. interacting South Asian x treatment, etc. Also, within all these interaction models, we will control for ethnicity as a categorical variable, which will account for difference by ethnicity at baseline.

Thus, we will have two approaches to incorporating ethnicity variables into our analysis: interactions with a binary variable (e.g. pupils identifying as Black, compared with pupils identifying as White), and controlling with a categorical variable (e.g. White, Black, South Asian, etc).

Imbalance at baseline

We will present descriptive statistics for baseline and endline pupil self-report responses, reporting continuous variables as means with standard deviations and categorical variables as counts with percentages for each trial arm. These will be reported by participant-level characteristics, as the analysis is being conducted at the individual level.

To assess the balance of participant characteristics between the treatment and control groups at baseline, we will perform regression analyses. Balance checks will be conducted within randomization blocks to identify potential biases in the randomisation process. Any findings will be discussed within the report.

Missing data

We will report the number of complete cases at both baseline and endline to provide a clear overview of data availability. To explore the extent and potential mechanisms of missingness, we will first conduct cross-tabulations between missingness indicators and key variables such as treatment group, demographic characteristics (e.g. gender, age, ethnicity), baseline scores, survey mode (paper vs digital), and cluster. This will help to identify whether missingness is systematically associated with observed characteristics.

We will then fit a logistic regression model ('drop-out model'), where the dependent variable is a binary indicator for missingness (1 = missing, 0 = observed). The model will include the following covariates:

- Treatment assignment
- Baseline demographic variables (e.g. gender, age, ethnicity)
- Baseline scores
- Survey mode (paper or digital)
- Cluster indicator

If any covariates are statistically significant at the 5% level, this suggests the data are not Missing Completely At Random (MCAR). In particular, if the treatment assignment variable is significant, this may indicate Missingness Experimentally Not At Random (MENAR), where missingness is related to treatment effects.

Criteria for Imputation and Sensitivity Analyses

Imputation will be considered when:

- Missingness exceeds 5% for key outcomes or covariates
- The MCAR assumption is rejected, based on the results of the drop-out model

If data appear MCAR and missingness is <5%, we will use listwise deletion. If missingness is more substantial, we will use Multiple Imputation by Chained Equations (MICE) to preserve statistical power.

If the data are judged to be Missing At Random (MAR), we will proceed with MICE. The imputation model will include:

- All covariates predictive of missingness or the outcome
- Treatment assignment
- Survey mode
- Cluster indicators

In line with the YEF Analysis Guidance (Youth Endowment Fund, 2021), results from multiple imputation (MI) analyses will be reported alongside the headline impact estimates. The implications of the MI analysis will also be clearly discussed in the final report

We will perform at least 20 imputations, in line with standard guidance.

Handling Potential MNAR (MENAR) Scenarios

Where MENAR is suspected (e.g. when treatment assignment significantly predicts missingness), we will:

- Conduct imputation within treatment groups
- Perform sensitivity analyses using:
    - Baseline Observation Carried Forward (BOCF)
    - Control Drifted Observation Carried Forward (CDOCF), where drift is estimated via autoregression models in the control group

These approaches will allow us to assess how assumptions about the missing data mechanism affect treatment effect estimates.

Fidelity

Fidelity will be measured at the practitioner level, i.e. are the lessons taught in the way they are intended to be taught by PSHE teachers and police. Given the size of the trial and the number of PSHE teachers involved, classroom observations will not be adequate for ensuring fidelity in these cases. Going into the trial, we will have to rely on our school-based contacts and PSHE leads to ensure rollout of the intervention. After the trial concludes, we will send our school and police officer contacts a final IPE survey, which captures what was delivered and to whom.

Compliance and contamination will be measured using a question from the pupils endline surveys in which pupils report if they received a PSHE lesson taught by a police officer within the term. We acknowledge, however, that pupil self-reports of receiving the treatment will be imperfect due to recall error, misinterpretation, and endline attrition, and these data will therefore be interpreted with caution. The primary analysis will follow an intention-to-treat (ITT) approach, estimating the impact of being allocated to the intervention group regardless of actual uptake.

We will additionally perform supplementary analyses to estimate the complier average causal effect (CACE) using a two-stage least squares (2SLS) method. In the first stage, we will regress actual treatment receipt (whether a student attended the PSHE lesson) on treatment assignment (whether the school/year was randomized to receive the intervention). Treatment assignment will serve as the instrumental variable in this stage. In the second stage, we will regress the primary outcome on the predicted treatment receipt from the first stage. This will estimate the effect of the intervention for students who actually received the PSHE lesson. In both stages, we will cluster standard errors at the school-year-time-period level. These findings will be presented and interpreted in the final report.

Intra-cluster correlations (ICCs)

We will model the Intracluster Correlation (ICC) at two levels: year group clusters and Black students within year groups.

1. ICC for Year Group Clusters:

The ICC for year group clusters will be calculated at the school-year level, where the year group is treated as the relevant clustering unit. We assume an ICC of 0.2 for year group clusters.

2. ICC for Black Students within Year Groups:

For the second model, we focus on Black students within each year group, treating this subgroup as the relevant cluster. We hypothesize that outcomes for Black students may exhibit stronger within-group correlations than for the overall year group. Consequently, we assume an ICC of 0.25 for Black students, reflecting a higher level of clustering within this group compared to the entire year group.

Both ICCs will be computed in two stages:

1. Empty Model (Random Intercept Model):
    a. This model will have no covariates and will estimate the variance at the cluster level (either school-year for year group clusters, or Black student subgroups within year groups).
    b. The model is specified as a random intercept model, where we estimate the variance between clusters (year groups or Black student subgroups) and the residual variance within clusters.
2. Primary Analysis Model:
    a. In the primary analysis, we will incorporate relevant covariates, including demographic variables.

Data Reporting Plan

We will provide a summary of means and standard deviations for all continuous baseline and outcome measures. In addition, histograms will be used to illustrate the distribution of baseline and outcome data.

For categorical data, we will report counts (numerators and denominators) and corresponding percentages for each category.

Presentation of outcomes

We will use a multi-level regression model to estimate the effect size for all outcomes. The effect sizes will be standardized using Cohen's D, with the mean difference between the two groups as the numerator, and the pooled standard deviation (calculated using the variances of the treatment and control groups) as the denominator. To account for the ICC, the effect sizes will be adjusted using the design effect.

To reflect the uncertainty around the estimated effect size, we will calculate and report 95% confidence intervals (CIs) for all estimates. For Cohen's D, the confidence interval can be derived using the standard error of the effect size. Effect sizes will be calculated using total variance to account for the nested data structure and potential differences between clusters. Confidence intervals will be adjusted for the clustering effect using robust standard errors.

References

Abadie, A, Athey, S, Imbens, GW, Wooldridge, JM. (2023). When Should You Adjust Standard Errors for Clustering?. *The Quarterly Journal of Economics,* 138, 1. https://doi.org/10.1093/qje/qjac038

Aspinall, P.J. (2020). Ethnic/Racial Terminology as a Form of Representation: A Critical Review of the Lexicon of Collective and Specific Terms in Use in Britain. *Genealogy*, *4*, 87. https://doi.org/10.3390/genealogy4030087

Collins, J, Horton, K, Gale-St. Ives, E et al. (2023). A Systematic Review of Autistic People and the Criminal Justice System: An Update of King and Murphy (2014). *J Autism Dev Disord* 53, 3151–3179. https://doi.org/10.1007/s10803-022-05590-3.

Goodman R, Meltzer H, Bailey V (1998) The Strengths and Difficulties Questionnaire: A pilot study on the validity of the self-report version. European Child and Adolescent Psychiatry, 7, 125-130.

Kogachi, K., & Graham, S. (2021). Same-race friendship preference across the middle school years: The role of school racial/ethnic context. *Developmental Psychology, 57*(12), 2134–2149. https://doi.org/10.1037/dev0001263

Murray, D. M., Short, B. J., & Tait, M. A. (1996). Intraclass correlation estimates in a school-based smoking prevention study. *American Journal of Epidemiology, 144*(4), 425–433. https://doi.org/10.1093/oxfordjournals.aje.a008938Oxford Academic

Nadal, K. L., & Davidoff, K. C. (2015). Perceptions of Police Scale (POPS): Measuring Attitudes towards Law Enforcement and Beliefs about Police Bias. *Journal of Psychology and Behavioral Science*, *3*(2). https://doi.org/10.15640/jpbs.v3n2a1

Parker, K., Nunns, M., Xiao, Z., Ford, T., Stallard, P., Kuyken, W., Axford, N., & Ukoumunne, O. C. (2025). Patterns of intra-cluster correlation coefficients in school-based cluster randomised controlled trials of interventions for improving social-emotional functioning outcomes in pupils: A secondary data analysis of five UK-based studies. *BMC Medical Research Methodology, 25*, 120. https://doi.org/10.1186/s12874-025-02574-6

Patel, C. (2020). Tackling police racism. *The World Today*, *76*(4), 13–15. Available at: https://www.jstor.org/stable/48750979 (Accessed: 16 August 2024).

Runnymede Trust. *The racialised harm of police strip searches: A response from the Runnymede Trust to a Home Office consultation.* (2024).

https://www.runnymedetrust.org//publications/the-racialised-harm-of-police-strip-searches-a-response-from-the-runnymede-trust-to-a-home-office-consultation

Sanders, M., Mitchell, C., & Ni Chonaire, A. (2020). Effect Sizes in Education Trials in England (SSRN Scholarly Paper 3532325). https://doi.org/10.2139/ssrn.3532325

Sanders, M., Ellingwood, J., Westlake, D., Bennett, V., Ewanich, K., Harrop, I., Ablitt, J., Corliss, C., & Hamlyn, A. (2024a). *Police in classrooms randomised controlled trial (efficacy): Evaluation protocol*. The Policy Institute at King's College London; Children's Social Care Research and Development Centre (CASCADE) at Cardiff University. Youth Endowment Fund. https://youthendowmentfund.org.uk/wp-content/uploads/2024/01/Police-in-Classrooms-Evaluation-Protocol-Efficacy.pdf

Sanders, M., Westlake, D., Ellingwood, J., Bancroft, K., & Bennett, V. (2024b). *Police in classrooms randomised trial – Pilot: Evaluation protocol*. King's College London; Cardiff University. Youth Endowment Fund. https://youthendowmentfund.org.uk/wp-content/uploads/2024/01/PiCl-Pilot-trial-protocol-Jan-2024.pdf

Selvarajah, S., Deivanayagam, T. A., Lasco, G., Scafe, S., White, A., Zembe-Mkabile, W., & Devakumar, D. (2020). Categorisation and Minoritisation. BMJ Global Health, 5(12), Article e004508. https://doi.org/10.1136/bmjgh-2020-004508

Thornberry, T. P., Lizotte, A. J., Krohn, M. D., Farnworth, M., & Jang, S. J. (1994). Delinquent Peers, Beliefs, and Delinquent Behavior: A Longitudinal Test of Interactional Theory. Criminology, 32(1), 47–83. https://doi.org/10.1111/j.1745-9125.1994.tb01146.x

Yesufu, S. (2013). Discriminatory Use of Police Stop-and-Search Powers in London, UK. *International Journal of Police Science & Management*, *15*(4), 281-293. https://doi.org/10.1350/ijps.2013.15.4.318

Young, SJ, Adamou, M, Bolea, B et al. (2011). The identification and management of ADHD offenders within the criminal justice system: a consensus statement from the UK Adult ADHD Network and criminal justice agencies. *BMC Psychiatry* 11, 32. https://doi.org/10.1186/1471-244X-11-32

Youth Endowment Fund. (2021). *Analysis guidance*. Youth Endowment Fund. https://res.cloudinary.com/yef/images/v1623145483/cdn/6.-YEF-Analysis-Guidance/6.-YEF-Analysis-Guidance.pdf

Youth Endowment Fund. (2025). *Racial disproportionality in violence affecting children and young people*. Available at: https://youthendowmentfund.org.uk/wp-

content/uploads/2025/02/YEF_Racial_Disproportionality_FINAL.pdf (Accessed 20th March 2025).

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413