STATISTICAL ANALYSIS PLAN

A cluster randomised controlled trial of a whole school trauma-informed practice programme in secondary schools

Anna Freud

Principal investigator: Julian Edbrooke-Childs

YOUTH ENDOWMENT FUND

A cluster randomised controlled trial of a whole school trauma-informed practice programme in secondary schools

Statistical analysis plan

Evaluating institution: Anna Freud

Principal investigator(s): Julian Edbrooke-Childs

Project title	A cluster randomised controlled trial of a whole school trauma-informed practice programme in secondary schools
Developer (Institution)	Knowledge Change Action and Warren Larkin Associations Ltd
Evaluator (Institution)	Anna Freud
Principal investigator(s)	Julian Edbrooke-Childs
SAP author(s)	Julian Edbrooke-Childs, Emily Goodacre, Angelika Labno, Suzet Tanya Lereya, Navya Malik, Sam Norton, Emily Stapley, Jessica Stepanous, Laura Talbot, and Jessica Deighton
Trial design	Two-arm cluster randomised controlled trial, with a nested mixed methods convergence design qualitative-driven implementation and process evaluation
Trial type	Efficacy
Evaluation setting	Mainstream secondary schools
Target group	The programme is a whole school approach but the target group for the evaluation is Year 8 and Year 9 pupils (aged 12-14 years) and school staff
Number of participants	Target number of pupils reached by the programme = 35,000-47,818 pupils across year groups in 50 schools

	Optimal sample size for the evaluation = 18,649 Year 8 and 9 pupils in 100 schools Optimal sample size for the evaluation (school staff) = 15,867
Primary outcome and data source	Externalising difficulties measured with the sum of the conduct problems and hyperactivity subscales of the Strengths and Difficulties Questionnaire (Goodman, Meltzer, & Bailey, 1998)
Secondary outcome and data source	**Pupil-reported surveys** Measure A: Strengths and Difficulties Questionnaire (Goodman et al., 1998) Outcomes: mental health difficulties, prosocial behaviour Variables: 1) internalising difficulties, 2) impact score, 3) prosocial behaviour Measure B: Student Engagement Instrument (Appleton et al., 2006) Outcomes: safe social connection(s) with teachers, safe social connection(s) with peers, school inclusion Variables: 4) teacher-student relationships, 5) peer support for learning, 6) control and relevance of school work, 7) future aspirations and goals Measure C: Illinois Bully Scale (Espelage & Holt, 2001) Outcome: bullying Variable: 8) bullying perpetration, 9) bullying victimisation **Staff-reported surveys** Measure A: Attitudes Related to Trauma-Informed Care (ARTIC 35) (Baker et al., 2021) Outcomes: knowledge and awareness, confidence, emotionally safe environment, vicarious trauma, empathy-focussed behaviours

	Variables: 1) underlying causes of problem behaviour and symptoms, 2) self-efficacy at work, 3) response to problem behaviour, 4) reactions to work, 5) empathy and control Measure B: Professional Quality of Life Scale (ProQOL) (Stamm, 2010) Outcomes: wellbeing, burnout Variables: 6) compassion satisfaction, 7) compassion fatigue, 8) burnout **Local school data (and National Pupil Database for longer-term outcomes)** Outcome: school attendance and exclusions Variables: 1) attendance, 2) exclusions, 3) permanent exclusions

SAP version history

Version	Date	Changes made and reason for revision
1.1	18th June 2025	In the original protocol, it was intended to include midline data in the primary analysis. However, since then, the prioritisation has shifted to conducting complete case analysis on baseline and endline data. A model that incorporates midline data will be used in robustness checks. The analysis section has been re-drafted in general to ensure alignment with YEF analysis guidance.
1.0 [*original*]		

Table of contents

Introduction

More Good Days At School (MGDAS) is a two-arm cluster randomised controlled trial. The trial is an evaluation of a whole-school trauma-informed practice programme, which focuses on training and capacity building for all school staff in understanding the impact of trauma, the importance of staff-pupil relationships, relational approaches, and strategies to look after one's own mental health.

The school is the unit of randomisation: half of the schools were randomised to receive the training, and half were to continue with existing wellbeing support practices at the school. The evaluation is focused on year 8 pupils, year 9 pupils and school staff. Schools were randomised in two cohorts: cohort 1 in November 2023 and cohort 2 in February 2024. Randomisation was stratified by Local Authority (LA) or Combined Authority (CA), with a minimum of two schools needed for LA/CA to be included in the randomisation.

There are three data collection periods: baseline, mid-programme (approximately 5 months after randomisation), and end of programme (after training is complete: approximately 16 months after randomisation). Data collection periods were coordinated between cohorts so that they occurred at the same stages for training inputs.

The primary outcome measure is externalising difficulties, as measured by the sum score of the conduct problems and hyperactivity subscales of the Strengths and Difficulties Questionnaire (Goodman et al., 1998). Secondary outcome measures are related to pupil and teacher wellbeing and school connections; the variables and measures are listed in detail in the Design Overview.

The main purpose of the analysis is to evaluate the impact of whole-school trauma-informed practice training on primary and secondary outcome measures for staff and pupils. Additional analyses will be conducted to test for differential effects of whole-school trauma-informed practice training, such whether the effect of training differs by ethnicity, free school meal (FSM) eligibility, special educational needs and disabilities (SEND) or Education, Health and Care Plan (EHCP) status, and staff-reported empathy and control. In addition, analyses will be conducted to clarify the mechanisms by which training impacts pupil outcomes, such whether safe social connections with teachers mediates the relationship between training and pupil's externalising difficulties.

Design overview

Trial design, including number of arms		Two-arm cluster randomised controlled trial
Unit of randomisation		Cluster (school)
Stratification variables (if applicable)		Local Authority or Combined Authority
Primary outcome	variable	Externalising difficulties
	measure (instrument, scale, source)	Sum of the conduct problems and hyperactivity subscales of the Strengths and Difficulties Questionnaire (Goodman et al., 1998)
Secondary outcome(s)	variable(s)	**Pupil-reported surveys** Variables: 1) internalising difficulties, 2) impact score, 3) prosocial behaviour Variables: 4) teacher-student relationships, 5) peer support for learning, 6) control and relevance of school work, 7) future aspirations and goals Variable: 8) bullying perpetration, 9) bullying victimisation **Staff-reported surveys** Variables: 1) underlying causes of problem behaviour and symptoms, 2) self-efficacy at work, 3) response to problem behaviour, 4) reactions to work, 5) empathy and control Variables: 6) compassion satisfaction, 7) compassion fatigue, 8) burnout **Local school data**

		Variables: 1) attendance, 2) fixed-term exclusions, 3) permanent exclusions
	measure(s) (instrument, scale, source)	**Pupil-reported surveys** ***Measure A:*** Strengths and Difficulties Questionnaire (Goodman et al., 1998) Outcomes: mental health difficulties, prosocial behaviour Variables: 1) internalising difficulties, 2) impact score, 3) prosocial behaviour ***Measure B:*** Student Engagement Instrument (Appleton et al., 2006) Outcomes: safe social connection(s) with teachers, safe social connection(s) with peers, school inclusion Variables: 4) teacher-student relationships, 5) peer support for learning, 6) control and relevance of school work, 7) future aspirations and goals ***Measure C:*** Illinois Bully Scale (Espelage & Holt, 2001) Outcome: bullying Variable: 8) bullying perpetration, 9) bullying victimisation **Staff-reported surveys** ***Measure A:*** Attitudes Related to Trauma-Informed Care (ARTIC 35) (Baker et al., 2021) Outcomes: knowledge and awareness, confidence, emotionally safe environment, vicarious trauma, empathy-focussed behaviours Variables: 1) underlying causes of problem behaviour and symptoms, 2) self-efficacy at work, 3)

		response to problem behaviour, 4) reactions to work, 5) empathy and control ***Measure B:*** Professional Quality of Life Scale (ProQOL) (Stamm, 2010) Outcomes: wellbeing, burnout Variables: 6) compassion satisfaction, 7) compassion fatigue, 8) burnout **Local school data** Outcome: academic attendance and exclusions Variables: 1) attendance, 2) fixed-term exclusions, 3) permanent exclusions
Baseline for primary outcome	**variable**	Externalising difficulties
	measure (instrument, scale, source)	Conduct problems and hyperactivity subscales of the Strengths and Difficulties Questionnaire (Goodman et al., 1998)
Baseline for secondary outcome	**variable**	**Pupil-reported surveys** Variables: 1) internalising difficulties, 2) no baseline for impact score, 3) prosocial behaviour Variables: 4) teacher-student relationships, 5) peer support for learning, 6) control and relevance of school work, 7) future aspirations and goals Variable: 8) bullying perpetration, 9) bullying victimisation **Staff-reported surveys** Variables: 1) underlying causes of problem behaviour and symptoms, 2) self-efficacy at work, 3) response to problem behaviour, 4) reactions to work, 5) empathy and control

		Variables: 6) compassion satisfaction, 7) compassion fatigue, 8) burnout **Local school data** Variables: 1) attendance, 2) fixed-term exclusions, 3) no baseline for permanent exclusions (given their relatively low frequency)
	measure (instrument, scale, source)	**Pupil-reported surveys** ***Measure A:*** Strengths and Difficulties Questionnaire (Goodman et al., 1998) Outcomes: mental health difficulties, prosocial behaviour Variables: 1) internalising difficulties, 2) no baseline for impact score, 3) prosocial behaviour ***Measure B:*** Student Engagement Instrument (Appleton et al., 2006) Outcomes: safe social connection(s) with teachers, safe social connection(s) with peers, school inclusion Variables: 4) teacher-student relationships, 5) peer support for learning, 6) control and relevance of school work, 7) future aspirations and goals ***Measure C:*** Illinois Bully Scale (Espelage & Holt, 2001) Outcome: bullying Variable: 8) bullying perpetration, 9) bullying victimisation **Staff-reported surveys** ***Measure A:*** Attitudes Related to Trauma-Informed Care (ARTIC 35) (Baker et al., 2021)

		Outcomes: knowledge and awareness, confidence, emotionally safe environment, vicarious trauma, empathy-focussed behaviours Variables: 1) underlying causes of problem behaviour and symptoms, 2) self-efficacy at work, 3) response to problem behaviour, 4) reactions to work, 5) empathy and control ***Measure B:*** Professional Quality of Life Scale (ProQOL) (Stamm, 2010) Outcomes: wellbeing, burnout Variables: 6) compassion satisfaction, 7) compassion fatigue, 8) burnout **Local school data** Outcome: academic attendance and exclusions Variables: 1) attendance, 2) fixed-term exclusions, 3) no baseline for permanent exclusions (given their relatively low frequency)

Sample size calculations overview

		Protocol	Randomisation
Minimum Detectable Effect Size (MDES)		Standardised mean difference = 0.15	Standardised mean difference = 0.17
Pre-test/ post-test correlations	level 1 (participant)	-	-
	level 2 (cluster)	-	-
Intracluster correlations (ICCs)	level 1 (participant)	0.03	0.041
	level 3 (cluster)	-	-
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided
Average cluster size		186 pupils 159 staff	220 pupils 26 staff
Number of clusters	intervention	29	31
	control	29	31
	total	58	62

		Protocol	Randomisation
Number of participants	intervention	4650 pupils 3975 staff	6429 pupils 724 staff
	control	4650 pupils 3975 staff	7194 pupils 896 staff
	total	9300 pupils 7950 staff	13623 pupils 1620 staff

The cluster sample size was determined a priori using a minimum detectable effect size (MDES) of 0.15 (standardised mean difference). The protocol presented different scenarios for sample size calculations based on a MDES between 0.10 and 0.20 (see Table 4 in the trial protocol). Presented in the table above is the sample size calculation for MDES of 0.15, which is in line with a projected school recruitment of 55 schools inflated to account for 15% attrition (58 in total). The estimate for the mean and SD of externalising difficulties in the control arm was based on the combined female and male scores from Elia et al. (2020). The intraclass correlation coefficient was assumed as 0.03. There is a lack of evidence on which to base assumptions for the pre-post correlation of externalising difficulties, so the sample size calculations were not adjusted to account for this. STATA was used for the power calculation using the following code: (MDE 0.15) power twomeans 5.21 4.76, sd(2.96) rho(0.03) m1(186) m2(186) power(0.8) cvcluster(0.5) cluster = 25 schools and 4,650 pupils per arm (increased to 29 schools per arm based on 15% school attrition).

The primary population of interest is pupils; estimates are also presented for staff. The pupil and staff sample sizes were determined from figures from three initial Local Authorities (LAs; Kent, Slough and Wiltshire). From this, 186 pupils per school were assumed after accounting for 35% study refusal and 25% attrition, resulting in an estimated sample size of 9300 pupils in total. 159 staff per schools were assumed after accounting for 35% study refusal and 25% attrition, resulting in 7950 staff in total.

The section with actual sample size and MDES at randomisation were updated following randomisation in February 2024. The average cluster size for pupils who completed the

primary outcome (SDQ externalising symptoms items) was 219.73, rounded to 220 pupils in the table. Furthermore, intracluster correlation was calculated as 0.041 and coefficient of variation was 0.366. With these parameters, we are powered to detect MDES of 0.17. Please see the following STATA code, which estimates the number of schools as 24.3 and the number of pupils per arm as 5333.02, inflated to 29 schools and 6372 pupils per arm based on 15% school attrition: power twomeans 5.21 4.7, sd(2.96) rho(0.041) m1(219.73) m2(219.73) power(0.8) cvcluster(0.366) nfrac display "schools per arm = " ceil(r(K1)/.85) display "pupils per arm = " round(ceil(r(K1)/.85)*219.73).

Analysis

Analyses will be conducted by an analyst masked to intervention group assignment following the intention-to-treat principle. The hypotheses and general analysis framework were decided a priori; the first draft analysis plan was written and approved during the period of data collection after recruitment but before baseline data collection was completed. The analysis plan was updated after the midline data collection, but before endline data collection was completed. Analyses other than the primary analysis are exploratory in nature. The following analyses will be conducted:

- Primary analysis
    - Complete case analysis using baseline and endline data for the primary outcome.
- Secondary analysis
    - Complete case analysis with secondary pupil and staff outcomes.
- Exploratory subgroup analysis
    - Moderation analyses by ethnicity, free school meal eligibility (FSM), special educational needs (SEN) status using the primary analysis model.
- Missing data analysis and robustness check of the primary analysis
    - If data are missing at random and:
        - If primary outcome is missing conditional on covariates: primary analysis model to include covariates that predict missingness.
        - If covariates are missing conditional on covariates/outcomes: multiple imputation and non-response weighting conducted.
- Tertiary analyses
    - Moderation analysis within the primary analysis
    - Mediation analysis within the primary analysis
    - Complete case analysis for tertiary outcomes
    - Longer-term outcome analysis
- Compliance analysis
    - Complier average causal effect (CACE) analysis using the primary outcome.
- Implementation analysis
    - Descriptive analysis of changes to policies and practices according to assigned condition.

Standard data quality checks will be performed in the data preparation stage before analysis, including impossible values, distribution of scores and residuals, outliers, low frequency categorical variables, linearity, and homogeneity of variance. Violations will be managed with

appropriate techniques. Furthermore, we will explore the amount of missingness in the data and the missingness mechanism to determine whether data are missing at random before proceeding with subsequent missing data analyses. Descriptive statistics of the implementation data will be used to identify levels of implementation and engagement.

There were no interim analyses for efficacy.

This project contains multiple secondary and tertiary outcomes and analyses. As a result, there is a higher risk of false positives overall. To mitigate this, false discovery rate (FDR) multiple comparison correction using the Benjamini-Hochberg procedure will be applied to the significance level for tests for outcomes that are within the same measure, as these are likely to be correlated both theoretically and statistically. For example, FDR correction will be applied to the three SDQ secondary pupil outcomes (internalising, prosocial, impact scores). The R function *p.*adjust with FDR correction specified will be used to adjust the three p-values from these models. In a similar vein, FDR correction will be applied to outcomes within the same measure for tertiary and subgroup analyses, if applicable. Some of these analyses will be underpowered, increasing the chance of false negatives in some cases. All exploratory analyses must therefore be treated with caution and not interpreted individually.

Imbalance at baseline

Table 1 provides baseline descriptive statistics for intervention and control groups, as well as the p-values for the t-tests for continuous data and chi-square tests (post-hoc test p-values in brackets) for count data. Members of the research team who are unblinded to intervention group allocation have access to the data in this table.

The school-level percentage of pupils eligible for free school meals and percentage of pupils with SEND or EHCP status was acquired through public data and averaged for intervention and control schools.

The pupil and staff descriptive statistics are from the baseline self-report pupil and staff surveys. Data from participants who had complete primary outcome data available (SDQ externalising for pupils and ARTIC-35 for staff) are included in the table. Please note that data on year group (year 8 or year 9) was missing for cohort 1 schools at the time of this analysis. Imbalance will be investigated again – including with complete year group data – when the endline data collection is complete.

In terms of imbalance at baseline, there were more females and less males in the intervention schools compared to control schools. This can be explained by a higher proportion of boys' schools present in the control group. Furthermore, intervention schools had a higher mean score on three out of the five ARTIC-35 sub-scales (Response to Problem Behaviour, On the Job Behaviour (Empathy and Control) and Reactions to Work) compared to control schools. This indicates that staff in the intervention schools have greater alignment with trauma-informed practice beliefs in those areas compared to the control schools. This may be partially explained by the differences in the baseline staff response rate between intervention and control schools (intervention: 35.1%, n = 29 schools; control: 28.3%, n = 28 schools). Please note that the response rates do not include schools who have not yet provided total school staff numbers; two are missing in the intervention arm, and three in the control arm. Again, this will be reviewed when endline data collection is complete.

Table 1. Baseline descriptive statistics of schools, pupils, and staff at baseline, split by intervention and control group.

	Intervention group (n = 31 schools)	**Control group (n = 31 schools)**	**p**
School-level	Mean (SD)	Mean (SD)	
Percentage of pupils with FSM eligibility	28.6% (10.6%)	28.2% (13.5%)	0.89
Percentage of pupils with SEND/EHCP status	15.6% (4.5%)	16.5% (4.6%)	0.41
Pupils	Mean (SD)	Mean (SD)	
SDQ Externalising Symptoms	7.58 (4.24), n = 6429	7.52 (4.22), n = 7194	0.41
Year group (Cohort 2 only)	Count (%)	Count (%)	0.25
Year 8	2659 (41.4%)	2822 (39.2%)	
Year 9	2518 (39.2%)	2796 (38.9%)	
Missing	1252 (19.5%)	1576 (21.9%)	
Gender			<.001
Female	3151 (49.0%)	3224 (44.8%)	(<.001)
Male	2978 (46.3%)	3668 (51.0%)	(<.001)
Non-binary	51 (0.8%)	64 (0.9%)	
Questioning	46 (0.7%)	52 (0.7%)	
Other	58 (0.9%)	59 (0.8%)	

	Intervention group (n = 31 schools)	Control group (n = 31 schools)	p
Prefer not to say	145 (2.3%)	127 (1.8%)	
Ethnicity			0.13
Asian/Asian British	1492 (23.2%)	1753 (24.4%)	
Black/African/ Caribbean/Black British	398 (6.2%)	426 (5.9%)	
Mixed/Multiple ethnic groups	267 (4.2%)	346 (4.8%)	
Other ethnic group	244 (3.8%)	270 (3.8%)	
White	3699 (57.5%)	4074 (56.6%)	
Prefer not to say	329 (5.1%)	325 (4.5%)	
Staff			
ARTIC-35 sub-scales	Mean (SD), n = 724	Mean (SD), n = 896	
Underlying Causes	4.75 (0.76)	4.67 (0.77)	0.04
Response to Problem Behaviour	5.05 (0.82)	4.90 (0.86)	<0.01
On Job Behaviour (Empathy and Control)	5.22 (0.75)	5.14 (0.74)	0.02
Self-Efficacy at Work	5.30 (0.92)	5.25 (0.94)	0.21
Reactions to Work	5.06 (0.78)	4.96 (0.84)	0.02
Gender	Count (%)	Count (%)	0.36
Female	515 (71.1%)	610 (68.1%)	
Male	206 (28.5%)	280 (31.3%)	
Non-binary	0 (0%)	0 (0%)	
Questioning	1 (0.1%)	0 (0%)	
Other	0 (0%)	1 (0.1%)	
Prefer not to say	2 (0.3%)	5 (0.6%)	
Ethnicity			0.26
Asian/Asian British	120 (16.6%)	160 (17.9%)	
Black/African/ Caribbean/Black British	24 (3.3%)	16 (1.8%)	
Mixed/Multiple ethnic groups	13 (1.8%)	17 (1.9%)	
Other ethnic group	20 (2.8%)	15 (1.7%)	
White	539 (74.4%)	677 (75.6%)	
Prefer not to say	8 (1.1%)	11 (1.2%)	

Note: ARTIC-35 = Attitudes Related to Trauma-Informed Care-35, EHCP = Education, Health and Care Plan, FSM = free school meals, SEND = special education needs and disabilities, SDQ = Strengths and Difficulties Questionnaire.

Primary analysis

The primary outcome is externalising difficulties measured using the sum of the conduct problems and hyperactivity subscales of the Strengths and Difficulties Questionnaire (Goodman et al., 1998). It will be included in the model as one total score (Goodman et al., 2010).

The primary analysis model to assess the intervention effect is a mixed effects model of complete cases at baseline and endline. This model will form the basis of the main trial report that will be submitted for publication. In the original protocol, it was intended to include midline data in the primary analysis. However, since then, the prioritisation has shifted to conducting complete case analysis on baseline and endline data. A model that incorporates midline data will be used in robustness checks, as reported in the *Missing data analysis and robustness check of the primary analysis* section.

Fixed effects will include group allocation (intervention/control), baseline externalising symptoms, and prognostic factors of cohort (cohort 1 (randomised November 2023) and cohort 2 (randomised February 2024)), gender (dummy coded: male/female/gender diverse, with the reference set as male) and Local Authority or Combined Authority (LA/CA; dummy coded).

Due to an initial imbalance in gender distribution between intervention and control groups at baseline (see Table 1), gender is included as a covariate in the analysis. Before conducting the analysis, group imbalance will be reassessed, considering school withdrawals that occurred post-randomisation. Any additional imbalances identified will be incorporated as covariates.

LA/CA was used as a stratifier in the randomisation and therefore it is intended to be included in the analysis as a fixed effect. However, there may be issues with model estimation due to a small number of schools per LA/CA (minimum of two). In this case, we will remove LA/CA as a fixed effect from the analysis.

Random intercepts for pupils/staff and schools will also be specified, depending on whether the outcome is pupil- or staff-specific. The model will have two levels, nesting pupil within school.

For pupil (i) in school (j), we define the model as:

$$(\text{endline externalising difficulties}_{ij}) = \beta_0 + \beta_1(\text{group allocation}_j) + \beta_2(\text{baseline externalising difficulties}_{ij}) + \beta_3(\text{cohort}_j) + \beta_4(\text{gender}_{ij}) + \beta_5(\text{LA/CA}_{ij}) + u_{0j} + v_{0i} + e_{ij}$$

For the mixed effects model, the package *lme4* version (version 1.1-35.1) in R will be used (Bates et al., 2015). Example syntax for the primary outcome of pupil externalising symptoms with LA/CA included as a fixed effect is:

lmer(endline_externalising ~ group + baseline_externalising + cohort_1_2 + male_female + male_diverse + LA_CA + (1 | school) + (1 | pupil), data = MGDAS_data)

Secondary analysis

Secondary analyses will be conducted on secondary pupil and staff outcomes. The model specification for the primary analysis - complete cases using baseline and endline data - will be used.

The secondary pupil outcomes are scores from the other subscales of the primary outcome measure (Strengths and Difficulties Questionnaire or SDQ; Goodman et al., 1998) and other outcomes theorised to be more directly impacted by intended changes to school staff behaviour arising from the intervention. False discovery rate (FDR) multiple comparison correction will be applied to the significance level for tests for tests within the same secondary outcome measure:

SDQ

- Internalising difficulties measured using the sum score of emotional symptoms and peer problems subscales of the SDQ.
- Impact of mental health difficulties measured using the impact score of the SDQ. Note that this model will not contain baseline impact score as a covariate as data were not collected at this time as per the protocol

- Prosocial behaviour measured using the prosocial behaviour subscale of the SDQ.

Teacher-student relationships

- Teacher-student relationships measured using the teacher-student relationships subscale of the Student Engagement Instrument (SEI; Appleton et al., 2006)

School exclusions

Logistic mixed-effects models will be used to analyse data on school exclusions.

- School exclusions using local school data, with baseline data defined as the 23/24 academic year period and endline data defined as the 24/25 academic year period. Although school exclusions are count data, the numbers in each category are likely to be small and binary variables will be created (any exclusion vs. no exclusion).
- Permanent school exclusions using local school data, with baseline data defined as the 23/24 academic year period and endline data defined as the 24/25 academic year period. Binary variables will be created (yes/no).

Example syntax is presented below:

```
glmer(endline_excluded ~ group + baseline_excluded + cohort_1_2 + male_female + male_diverse + LA_CA + (1 | school) + (1 | pupil), data = MGDAS_data, family = binomial)
```

Staff secondary outcome analysis will follow the same specification as pupil outcomes:

ARTIC 35

- Underlying causes of problem behaviour and symptoms measured using the underlying causes of problem behaviour and symptoms subscale of the Attitudes Related to Trauma-Informed Care (ARTIC 35; Baker et al., 2021)
- Self-efficacy at work measured using the self-efficacy at work subscale of the ARTIC 35
- Response to problem behaviour measured using the response to problem behaviour subscale of the ARTIC 35

- Reactions to work measured using the reactions to work subscale of the ARTIC 35
- Empathy and control measured using the empathy and control subscale of the ARTIC 35

ProQOL

- Compassion satisfaction measured using the compassion satisfaction subscale of the Professional Quality of Life Scale (ProQOL; Stamm 2010)
- Compassion fatigue measured using the compassion fatigue subscale of the ProQOL.

Exploratory subgroup analysis

Rather than stratifying the analysis by demographic subgroups, the intention is to extend the primary analysis by including interaction terms between the group allocation variable and each demographic variable listed below. The rationale for this analysis is to explore whether the intervention effect varies across key demographic subgroups. tests for variables with multiple categories will be reported at each level with 95% confidence intervals (CIs) and with the inference based on an omnibus test across all levels.

- Ethnicity will be derived from self-reported pupil and staff survey data. Response options include broad ethnic classifications: Asian, Black, Mixed, White, Other, and Prefer not to say. These categories will be dummy coded with White as the reference category. They will then be integrated as interaction terms with group allocation.
- Pupil FSM eligibility will be obtained from local school administrative data. This will be dummy coded (yes/no) and will be included as an interaction term with group allocation.
- Pupil SEND/EHCP status will be obtained from local school administrative data. SEND/EHCP status will be a binary category, and pupils receiving any SEND support or who have an EHCP will be counted within the SEND/EHCP group. This will be dummy coded (yes/no) and will be included as an interaction term with group.

Missing data analysis and robustness checks of the primary analysis

Missing data analysis will be conducted to clarify the missing data mechanism and support valid inferences. Intervention schools that have withdrawn post-randomisation will still be asked to participate in the follow-up data collection periods in an attempt to minimise missing data. We will explore the amount of missing data in the primary outcome and covariates. To

assess whether data are missing completely at random (MCAR), Little's MCAR test will be conducted using the LittleMCAR() function in the *mice* package. If data are MCAR, the complete case analysis may provide unbiased estimates.If data are more than 5% missing, we will evaluate the extent to which data are systematically missing. This will be achieved through multi-level logistic regression, which will be used to identify variables that predict missingness for the primary outcome and covariates (0 = not missing, 1 = missing) in separate models.

Missing covariate data

If covariates (i.e. baseline externalising symptoms) are missing conditional on other covariates or outcomes (missing at random (MAR)), multiple imputation (MI) will be conducted, and the results will be compared with the complete case analysis.

Multiple imputation will be conducted for covariates used in the primary analysis. The package *mice* will be used to impute missing values (van Buuren & Groothuis-Oudshoorn, 2011). When building the predictor matrix, variables that strongly correlate with missingness will be included as predictors, and the temporal order of variables will be maintained (i.e. endline externalising will not be used as a predictor of baseline externalising symptoms). School will also be specified as a predictor in the predictor matrix to account for the clustered nature of the data. The number of imputations will match the percentage of missing data. Analyses will be run with these datasets with the pooled estimates derived. If there are convergence issues with using multiply imputed data, full information maximum likelihood will be used instead.

Missing primary outcome data

If the primary outcome is missing conditional on covariates (MAR), then the primary analysis (complete cases analysis) will include covariates that predict missingness into the model. The results will be compared with the complete case analysis that does not include these covariates.

Furthermore, non-response weighting will be applied to account for biases resulting from differential participation of groups in completing the primary outcome. Weights will be computed by first conducting a logistic regression to predict the probability of response of the primary outcome based on covariates and then calculating the inverse of the response

probability (Valliant et al., 2013). Weight will then be normalised and then included in the primary analysis model.

If, in the cases above, the results are not similar with the complete case analysis, the data are likely missing not at random (MNAR) and further sensitivity analysis is required. Two options are pattern mixture models or selection models: the guidance by Carpenter & Kenward (2007, p. 119) will be followed in this scenario.

Imbalance of staff data

To explore the impact of imbalance at baseline in staff-reported ARTIC-35 subscales (see Table 1), an additional model will be run that includes baseline school-level ARTIC-35 subscale scores as a control variable in the primary analysis with pupil externalising symptoms as the outcome variable. If any further staff-reported variables are imbalanced at baseline, these will also be included.

Midline SDQ externalising data

Furthermore, to factor in midline SDQ externalising scores into the analysis, an adjustment to the primary analysis will be made where full information maximum likelihood will be used to enable include of participants who have either midline or endline scores for the SDQ externalising symptoms. Results from this model will be compared to the primary analysis model. The model is as follows:

lmer(externalising ~ group * time + baseline_externalising + cohort_1_2 + male_female + male_diverse + LA_CA + (1 | school) + (1 | pupil), data = MGDAS_data)

Tertiary analyses

Moderation analysis

In the primary analysis model, a cross-level interaction between group allocation and scores on the staff-reported empathy and control sub-scale of the ARTIC 35 will be included to examine whether allocation to treatment vs. control groups predicts different levels of externalising difficulties for pupils in schools where staff report higher levels of empathy-focussed behaviours.

Longitudinal mediation analysis

A longitudinal mediation analysis will be conducted to test whether safe social connections with teachers measured by the SEI at midline mediates the relationship between group allocation and changes in externalising difficulties from baseline to endline, accounting for pupils nested within schools. For the mediation model, the package *lavaan* (version 0.6-17) in R will be used (Rosseel, 2012). This will be an exploratory analysis that may be used to drive future research.

Tertiary outcomes

Furthermore, complete case analysis will be conducted on the following tertiary outcomes:

Student Engagement Instrument (SEI)

- Peer support for learning measured using the peer support for learning subscale of the SEI
- Control and relevance of school work measured using the control and relevance of school work subscale of the SEI
- Future aspirations and goals measured using the future aspirations and goals subscale of the SEI

Illinois Bully Scale (IBS)

- Bullying perpetration measured using the bullying perpetration subscale of the Illinois Bully Scale (IBS; Espelage & Holt, 2001)
- Bullying victimisation measured using the bullying victimisation subscale of the IBS
- School attendance using local school data at mid-programme and end-of-programme

Longer-term outcomes

Long-term outcomes will be analysed once data are received from the National Pupil Database. After the end of the YEF-funded study (winter 2025), Anna Freud will link the quantitative data on pupils to national data from the National Pupil Database through the Office for National Statistics Secure Research Service. This will enable us to examine longer-term impacts of the programme on attendance, exclusions, and permanent exclusions. As this

is beyond the timeframes of the present study, it will be conducted as non-costed work by Anna Freud, with the intention of submitting a peer-review journal article. Longer-term analyses will also be conducted by YEF as part of the YEF Data Archive.

- Longer-term school attendance will be measured using data from the National Pupil Database after winter 2025
- Longer-term school exclusions will be measured using data from the National Pupil Database after winter 2025
- Longer-term permanent school exclusions will be measured using data from the National Pupil Database after winter 2025
- Academic attainment will be measured using Key Stage 4 attainment after winter 2025.

Compliance analysis

Compliance will be measured using activity data collected by the MGDAS programme team, for example, number and types of sessions delivered and number and types of school staff attending each session. The average percentage of eligible staff attendance across the three types of sessions (whole school staff, pastoral and inclusion leads, and senior leadership team) will be used as the indicator of compliance.

Complier average causal effect (CACE) analysis will be used to estimate the intervention effect on the primary outcome for a hypothetical scenario where everyone was adherent using the primary analysis model (complete cases). The average percentage of eligible staff attendance across training sessions for each intervention school will be used as the compliance measure. The threshold for compliance will be specified as 80%. If there are too few schools reaching 80% compliance or above, we will use the upper quartile of the average percentage as the threshold. We will aim to conduct CACE analysis through bespoke programming due to the outcome data being at the pupil level but the compliance metric being at the cluster (school) level. This will account for the first stage regression at the cluster level and the second stage regression at the individual outcome level with adjustment for cluster. If bespoke programming is not possible, the caceCRTBoot function in the *eefAnalytics* package in R will be used to conduct the CACE analysis, accounting for clustering at the school level. CACE analysis will be conducted irrespective of whether the intervention effect is significant or non-significant. If the effect is non-significant, this sensitivity analysis will be useful to determine whether a non-significant main intervention effect was due to the intervention itself or a lack of adherence.

Implementation analysis

The support description survey completed by the single point of contact at each school, to examine changes to policies and practices implemented by schools after the MGDAS programme. We will create a composite score of school-level compliance with assigned condition based on the number and type of changes implemented and the stage of implementation. It should be noted that we are not attempting to restrict implementation of changes to policy and practice for schools allocated to business as usual only.

A descriptive analysis will be undertaken to examine whether there are differences in the levels of changes to policies and practices according to assigned condition (i.e., MGDAS and business as usual, business as usual only).

Interim analyses and stopping rules

Reports were prepared in March 2024 for the Data Monitoring and Ethics Committee (DMEC) covering recruitment and adverse events as part of monitoring of trial progress. However, no formal interim analysis considering either efficacy or futility will be undertaken. Stopping rules were not specified in the protocol and recruitment targets were met based on a priori power calculations. As a result, the only stopping rules will be where the DMEC makes a recommendation for the trial to terminate early based on safety concerns. We will continue to monitor withdrawals and assess study continuation from that. For example, a priori power calculations found that a minimum of 36 schools total is needed for a MDES of 0.20, which is the Youth Endowment Fund's MDES requirement.

Intra-cluster correlations (ICCs)

Intra-cluster correlations will be calculated pre- and post-test for both individuals (pupils and staff) and schools. ICCs will be reported for all models with 95% confidence intervals.

Presentation of outcomes

Standardised mean difference will be used to calculate the effect size. Total variance will be used in the calculation of the effect size, to account for the nested structure of the data.

The exact specification of the numerator and denominator is detailed below:

- Numerator = Adjusted mean difference estimate at each time point from the primary intervention effect analysis described above
- Denominator = Pooled unconditional variance for both intervention and control groups.

Bootstrapped confidence intervals will be used to reflect statistical uncertainty.

References

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument, Journal of School Psychology, 44(5), p.427-445.

Baker, C. N., Brown, S. M., Overstreet, S., Wilcox, P. D., & New Orleans Trauma Informed Schools Learning Collaborative, Validation of the Attitudes Related to Trauma-Informed Care Scale (ARCTIC). (2021). Psychological Trauma: Theory, Research, Practice, and Policy, 13(5), p.505-513.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), p.1-48. https://doi.org/10.18637/jss.v067.i01

Carpenter, J., & Kenward, M. (2007). Missing data in randomised controlled trials - a practical guide. London: LSHTM.

Elia, C., Karamanos, A., Silva, M. J., O'Connor, M., Lu, Y., Dregan, A., et al., Harding, S. (2020). Weight misperception and psychological symptoms from adolescence to young adulthood: longitudinal study of an ethnically diverse UK cohort. BMC Public Health, 20, p.712. https://doi.org/10.1186/s12889-020-08823-1Espelage, D. L., & Holt, M. K. (2001). Bullying and victimization during early adolescence: Peer influences and psychosocial correlates Journal of Emotional Abuse, 2:2-3, p.123-142.

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. Journal of Abnormal Child Psychology, 38, 1179-1191. https://doi.org/10.1007/s10802-010-9434-x

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(3), p.125-30.

Nelson, G. A. (2023). fishmethods: fishery science methods and models in R. https://cran.r-project.org/package=fishmethods

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), p.1-36. https://doi.org/10.18637/jss.v048.i02

Stamm, B. H. (2010). The Concise ProQOL Manual, 2nd Ed. Pocatello, ID: ProQOL.org

Valliant et al. (2013) *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer Science Business Media.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.
https://doi.org/10.18637/jss.v045.i03

YOUTH ENDOWMENT FUND

Impetus

Home Office

youthendowmentfund.org.uk

hello@youthendowmentfund.org.uk

@YouthEndowFund

The Youth Endowment Fund Charitable Trust

Registered Charity Number: 1185413