Evaluation of Whole School SEND Review: A cluster randomised controlled trial
Statistical Analysis Plan

Evaluator (institution): Manchester Metropolitan University
Principal investigator(s): Professor Stephen Morris and Professor Cathy Lewin

Template last updated: August 2019

PROJECT TITLE[1]	Evaluation of Whole School SEND Review: A cluster randomised controlled trial
DEVELOPER	nasen
EVALUATOR	Manchester Metropolitan University
PRINCIPAL INVESTIGATORS	Stephen Morris and Cathy Lewin
SAP AUTHORS	Stephen Morris, Cathy Lewin, Peter Hick and Andrew Smith
TRIAL DESIGN	Two-arm, pragmatic, cluster randomised controlled trial with random allocation of schools
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Age range during the study: 12/13-15/16 years, recruitment and enumeration during KS3 with outcomes at KS4
NUMBER OF SCHOOLS	156
NUMBER OF PUPILS	59,971 all pupils 9,054 SEND pupils
PRIMARY OUTCOME MEASURE AND SOURCE	Marks in GCSE English Language (Exam Boards via schools)
SECONDARY OUTCOME MEASURE AND SOURCE	• Marks in GCSE Mathematics (Exam Boards via schools) • Grade in GCSE English Language (school records) • Grade in GCSE Mathematics (school records) • Unauthorised absence in previous school year (School records) • Authorised absences in previous school year (School records) • Fixed term or permanent exclusion in previous school year (School records) • Student wellbeing score (Total difficulties score, student 11-17 self-completion SDQ)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]	TBC	*N/A*

Table of contents

Introduction

This statistical analysis plan describes the proposed analysis of data from a cluster randomised controlled trial (CRCT) designed to evaluate the effectiveness of Whole School SEND (WSS) Review.

WSS Review is a programme developed and delivered by nasen (https://nasen.org.uk/), who provide training and support to schools, and who work closely with school leadership and special needs professionals. The programme aims to help schools prioritise SEND provision through encouraging leadership teams to take ownership of and to support school development of SEND – ultimately with the aim of improving pupil outcomes. WSS Review is a whole school intervention that aspires to be constructive, collaborative and owned by the school (rather than an audit or inspection process). It seeks to draw on and support existing expertise and good practice within and across schools. The intervention is delivered to SENDCos (special educational needs coordinators) who are expected to oversee the Whole School SEND (WSS) Review within their own school and to develop and implement a SEND Development Plan, targeting areas for improvement. The WSS Review process aims to raise awareness and give SENDCos more status such that they can become agents of change. Their role should shift from one with a pastoral focus to one that drives change in both teaching and learning; it is this that sets apart WSS Review from other SEND-related interventions, and is at the heart of the 'innovation' that WSS Review represents. Further details of the intervention including its theory of change can be found in the published trial protocols (S. Morris et al., 2020, 2021).

Design overview

This is a pragmatic two-arm parallel CRCT with whole schools allocated at random to treatment and control conditions on a 1:1 basis (March 2021). The intervention is delivered to participating state secondary schools on a regional basis. To aid delivery, randomisation was stratified by region. The study population comprises pupils in trial schools entering Years 8 and 9 at September 2020 and within these focal cohorts, pupils with a SEND designation; that is with either an EHCP or "support".

The primary outcome is the mark obtained by pupils with a SEND designation in their GCSE English language examination to be sat in the summer of 2023 for Year 9 pupils, and 2024 for pupils entering Year 8 in September 2020. Examination marks will be obtained direct from schools by the Fisher Family Trust (FFT) on behalf of the evaluation team. As pupils will have obtained marks through sitting examinations from different awarding organisations, prior to analysis marks will be standardised.

Secondary outcomes for pupils designated SEND are:

- the standardised mark obtained in GCSE mathematics examination
- the grades obtained in GCSE mathematics and English language examinations
- number of unauthorised and authorised absences in the school year 2022/23 for Year 9 pupils and 2023/24 for 8 Year pupils
- number of temporary or permanent exclusions in the school year 2022/23 for Year 9 pupils and 2023/24 for Year 8 pupils, and
- the mean difficulties scores obtained from a Strengths and Difficulties Questionnaire (SDQ) completed June/July 2022 for Year 9 pupils and June/July 2023 for Year 8 pupils.

These data, with the exception of the SDQ, are collected direct from school information systems at baseline (autumn term 2020) by FFT, and then again in the Autumn terms of 2023 and 2024 for the purposes of measuring post-intervention outcomes. SDQ measures are obtained from self-completion questionnaires delivered on-line to pupils in school settings at baseline (April-June, 2021), at June/July 2022 for Year 9 pupils and June/July 2023 for Year 8 pupils.

Other secondary outcomes will be examined *for all pupils*, whether designated SEND or otherwise. These outcomes will be:

- the standardised mark in GCSE English language from summer of 2023 national examinations for pupils entering Year 9 in September 2020 and 2024 for pupils entering Year 8 in September 2020;
- the standardised mark in GCSE mathematics from summer of 2023 national examinations for Year 9 pupils and 2024 for Year 8 pupils; and
- the mean difficulties scores obtained from a pupil-self-completion SDQ administered in June/July 2022 for Year 9 pupils and June/July 2023 for Year 8 pupils.

Trial design, including number of arms		Two arm cluster randomized controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		Region
Primary outcome	variable	Mark obtained in GCSE English language
	measure (instrument, scale, source)	Standardised marks in GCSE English language obtained via schools from exam boards
Secondary outcome(s)	variable(s)	• Standardised Mark obtained in GCSE Mathematics • Grade obtained in GCSE English Language • Grade obtained in GCSE Mathematics • Unauthorised absences • Authorised absences • Exclusions (fixed term / permanent) from school • Total difficulties reported
	measure(s) (instrument, scale, source)	• Standardised Marks obtained from exam boards via schools • Grades recorded as 0-9, where 0 is an unclassified score at GCSE, obtained from schools (equivalent to results reported in NPD) • Count of authorised absences in the last full academic year – school records • Binary zero/one response – whether a least one unauthorised absence recorded in the last full academic year • Binary zero/one indicator – whether a least one exclusion recorded in the last full academic year • Total difficulties reported – child self-completion age 11-17 single-sided SDQ questionnaire.
Baseline for primary outcome	**variable**	Prior attainment in English reading at KS2
	measure (instrument, scale, source)	Score at KS2 test score in reading obtained from schools
Baseline for secondary outcome	**variable**	As appropriate: • Prior attainment in either Mathematics or English reading at KS2 • Count of authorised absences in the School Year prior to randomisation • Count of all absences in the School Year prior to randomisation • Total difficulties reported prior to commencement of the intervention

measure (instrument, scale, source)	• For attainment baseline measures these are raw continuous scores at KS2 obtained from schools • Absence measures obtained from school records – coded as counts (authorised or all absences) • Total difficulties obtained as a continuous measure derived from self-reports via SDQ student self-completion questionnaire for 11-17 year olds

As will be explained below, sample estimates of average effects will be obtained from separate regression models for each primary and secondary outcome. The primary outcome or one of the secondary outcomes will be the response or dependent variable. Sample estimates of treatment effects on the primary outcome will be adjusted through the inclusion of prior attainment in reading at KS2 as a covariate in the regression model. Other variables to be used as covariates in separate secondary outcome analysis, measured prior to randomisation, are: prior attainment in mathematics at KS2, absences in the school year prior to randomisation and total difficulties score prior to randomisation obtained from SDQs administered in the summer term of 2021 (April-June).

Sample size calculations overview

		Protocol			Randomisation		
		All pupils	**SEND pupils**	**FSM pupils**	**All pupils**	**SEND pupils**	**FSM pupils**
Minimum Detectable Effect Size (MDES)		0.19	0.20	0.19	0.19	0.21	0.20
Pre-test/ post-test correlations	level 1 (pupil)	0.70	0.70	0.70	0.70	0.70	0.70
	level 2 (class)	n/a	n/a	n/a	n/a	n/a	n/a
	level 3 (school)	0.32	0.32	0.32	0.32	0.32	0.32
Intracluster correlations (ICCs)	level 2 (class)	n/a	n/a	n/a	n/a	n/a	n/a
	level 3 (school)	0.20	0.20	0.20	0.20	0.20	0.20
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two	Two	Two
Average cluster size		180	25	42	163	17	31
Number of schools	intervention	80	80	80	78	78	78
	control	80	80	80	78	78	78
	total	160	160	160	156	156	156
Number of pupils	intervention	14,400	2,000	3,360	14,768	1,924	3,721
	control	14,400	2,000	3,360	14,934	2,257	4,091
	total	28,800	4,000	6,720	29,702	4,181	7,812

Sample size calculations made at the point in time at which the protocol for this trial was written are presented in the table above. Alongside these are similar calculations based on sample sizes at the time of randomisation. Assumptions for pre/post-test correlations remain the same in the two sets of calculations as does the assumed intraclass correlations. Type I and II error rates and the assumption of two-sided tests for statistical significance are also maintained. The 'as randomised' sample consisted of 156 schools; though 157 schools were randomised with one school subsequently withdrawing from the study. This is compared to the proposed protocol sample size of 160 schools.

It is important to note the average cluster sizes, which are the average number of pupils in a given year by school, were based on assumed or projected arithmetic mean in the protocol for all students as well as SEN and FSM pupils. The calculations based on the 'as randomised' sample use the harmonic mean of the actual average school size for all, SEN and FSM pupils. Use of the harmonic mean is recommended by the authors of programs for sample size determination such as PowerUp, in order to take account of variable cluster sizes (Dong & Maynard, 2013). As can be seen the use of the harmonic mean has little influence on our calculations. The following table reports the relevant arithmetic and harmonic means in the 'as randomised' sample, based on data for pupils in Year 9 (values for Year 8 pupils are very similar):

	Arithmetic Mean school/cluster size	Harmonic mean school/cluster size	Total sample size
All pupils	190.4	162.6	29,702
SEND	26.8	17.2	4,181
FSM	50.1	31.2	7,815

What the tables above show is that the average size of clusters (that is schools), in terms of the arithmetic mean, is slightly larger in the 'as randomised' sample compared to that anticipated in the protocol. Furthermore, control group schools are on average very slightly larger than intervention schools.

The justifications for the assumptions made in sample size calculations at both protocol and at randomisation are as follows:

- It is standard practice in EEF trials for Type I and II error rates to be set at five and 20 per cent respectively on the basis that this represents an appropriate balance between the risks of false positive and false negative inferences
- Randomisation to intervention and control on a 1:1 basis was chosen in order to maximise statistical power
- The intervention delivery team had the capacity to deliver the intervention to around 100 schools, which meant an upper limit of around 200 schools could be recruited to the study and randomised. However, previous experience suggested that it would be difficult to recruit 200 schools and maintain their participation (S. P. Morris et al., 2018), therefore a less demanding target of 160 schools was proposed at the protocol stage. As noted above, 156 schools, with one additional school leaving the study subsequent to randomisation, eventually agreed to take part and were randomised
- Estimates of the correlation between KS2 raw score for English and GCSE English language attainment were obtained from analysis provided by EEF (Education Endowment Foundation, 2013)
- The value for the intra-class correlation coefficient used in these calculations was set at 0.20; this is quite conservative but is the assumption generally adopted in the design of EEF trials

Given the sample sizes projected at the protocol stage, minimum detectable effect sizes of 0.19 were obtained for an estimate based on samples of all pupils, 0.20 for samples of SEND pupils and 0.19 for FSM subgroups. Based on the samples achieved at randomisation equivalent minimum detectable effect sizes were 0.19, 0.21 and 0.20.

In order to provide some sense of the consequences of attrition on these calculations, we repeated them assuming first that five schools of average size attrite from the sample and then 10 schools, again of average size in terms of their harmonic mean. We assume that attrition occurs proportionately in both arms of the trial. For estimates derived from the SEND only sample, an achieved sample of 151 instead of 156 schools has no implications for the MDES of 0.21 above. A loss of 10 schools sees the MDES of 0.21 rise very slightly to 0.22. Twenty or more schools will have to attrite from the sample before appreciable loss in statistical power is seen.

These calculations are made using the PowerUp program in R statistical software.

Analysis

The analysis will proceed on the basis of the principle of intention to treat (ITT), whereby pupils are identified in the analysis as members of the intervention or control group on the basis of the allocation of their school to intervention and control conditions at randomisation, regardless of whether the school subsequently takes part in the intervention or not. Where schools leave the study subsequent to randomisation and ask that their data are deleted, records for the relevant pupils will be removed from the sample file. Such loss to the sample could lead to bias in sample estimates. Further discussion of approaches to assessing the consequences of sample loss and possible strategies for missing data are discussed below.

Primary outcome analysis

The primary analysis seeks an estimate of the average effect of intention to treat (AITT), of the intervention, on marks obtained in English GCSE examinations for pupils designated SEND. Analysis will proceed in two stages. First data from the Year 9 cohort of SEND students will be the focus of attention. At a later stage, reported separately, if implementation is deemed successful, further primary outcome analysis as described here will be undertaken on the Year 8 sample. As such, we consider there to be a single primary outcome as the two analyses will be separated by some considerable period of time.

For each Year group analysis, a sample estimate of AITT will be obtained from a hierarchical linear model of the following form:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2(X_{ij} - \bar{X}_j) + \beta_3\left(\bar{X}_j - \bar{X}\right) + \beta_4 R_j + \vartheta_j + \varepsilon_{ij} \ldots [1]$$

Here Y_{ij} is the standardised mark obtained by SEND pupil i in school j in their English language GCSE. Marks are standardised as a result of pupils sitting examinations set by different awarding organisations. Each student's mark will be standardised on the basis of the sample mean and sample standard deviation of the relevant awarding body's mark distribution. The variable T_j will take the value one if the pupil is in a school randomised to the intervention, zero otherwise. The sample estimate of the parameter β_1 is the estimate of AITT. X_{ij} represents student j's points score in their KS2 English reading test. This measure of prior attainment is entered into the model at the pupil level through the inclusion of a covariate centred on the school mean for each pupil and through the inclusion of a covariate capturing the average prior attainment at the school level centred on the grand mean for the sample. R_j captures the region in which school j is located and is included to reflect the fact that randomisation was stratified by region. The terms ϑ_j and ε_{ij} are random effects at the school and pupil levels and are assumed to be distributed normally in the population with zero means, variances σ^2 and τ^2 respectively, and for these variance to be conditionally uncorrelated. The intra-class correlation coefficient, or rho, is therefore $\rho = \sigma^2/\sigma^2 + \tau^2$. Parameter estimates will obtained using maximum likelihood and the 'mixed' suite of commands in STATA v17 statistical software.

For the primary outcome, sensitivity analysis will comprise the estimation of three further models in order to assess the consequences of regression adjustment for the sample estimates. The first model will be a simple variance components that will provide an unconditional estimate of ρ. The second will be a hierarchical linear model containing only the covariates T_j and R_j. Estimates from this model are equivalent to difference in means by stratum, and when compared to estimates from the primary analysis, permit us to assess the consequences for our estimates of the inclusion of prior attainment as a covariate. Third, estimates from an extended regression model with additional covariates to the main primary outcome model will be presented. These additional covariates will be gender, month of birth and an indicator of whether the pupil concerned had ever qualified for free school meals. Covariates measured at the school level and obtained from the school census will also be included in this model. These covariates will be the proportion of the school roll in the year 2018/19 that qualified for free school meals, proportion of the school roll that were EAL in 2018/19, proportion of the school roll SEN in 2018/19, and average Attainment 8 scores for the school in the year 2018/19. The extended model will examine the consequences for sample estimates of the inclusion of these additional adjustment factors.

Inference will be performed through constructing frequentist 95 per cent confidence intervals, derived from heteroskedastic robust standard errors.

Further proposed sensitivity tests for the primary analysis in relation to missing data are discussed below.

Secondary outcome analysis

A wide range of further secondary/exploratory analysis is proposed. As with the primary analysis, these will be performed and reported separately for Year 9 and 8 cohorts. These secondary analyses will involve obtaining estimates from a series of regression models, described in the Table below and similar to that proposed for the primary analysis. For students designated SEND, the table below sets out the regression models that will form the basis of the secondary analysis

Dependent variable	Model	Intervention group indicator	Region indicator	Further covariates	interference
GCSE mathematics standardised mark	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) KS2 mathematics points score at pupil and school levels 2) Gender 3) FSM; and 4) Month of birth	Robust standard errors/95 per cent confidence interval
GCSE English Grade 1-9	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) KS2 reading points score at pupil and school levels 2) Gender 3) FSM 4) Month of birth	Robust standard errors/95 per cent confidence interval
GCSE mathematics Grade 1-9	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) KS2 mathematics points score at pupil and school levels 2) Gender 3) FSM 4) month of birth	Robust standard errors/95 per cent confidence interval

Count response - number of authorised absences – school year 2022/23	Hierarchical negative binomial model - random effects at school and pupil levels	Yes	Yes	1) absences in the school year 2019/20 at the pupil level 2) Gender 3) FSM 4) Month of birth	Robust standard errors/95 per cent confidence interval
Binary response – at least one unauthorised absence in the school year 2022/23	Hierarchical binary logistic regression – random effects at school and pupil levels	Yes	Yes	1) absences in the school year 2019/20 at the pupil level 2) Gender 3) FSM 4) Month of birth	Robust standard errors/95 per cent confidence interval
Binary response – at least one exclusion from school in the school year 2022/23	Hierarchical binary logistic regression – random effects at school and pupil levels	Yes	Yes	1) absences in the school year 2019/20 at the pupil level 2) Gender 3) FSM 4) Month of birth	Robust standard errors/95 per cent confidence interval
-Total difficulties score – Strengths and Difficulties Questionnaire	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) Total difficulties score measured at the baseline 2) Gender 3) FSM 4) Month of birth	Robust standard errors/95 per cent confidence interval

Results from these regression models will be reported as regression coefficient estimates in all cases and, in addition, in the case of linear models as effect sizes consistent with Hedge's g (Durlak, 2009), as relative risk ratios in the case of logistic regression models and in the case of negative binomial models incident rate ratios. As these are secondary analyses, other than basic assessment of model fit, no further sensitivity testing is proposed and all analysis will be performed on the completed cases file.

Secondary analysis involves the testing of multiple hypotheses and therefore the risk of inflated Type I statistical errors where results are considered together. To take account of the family-wise error rate, the Holm-Sidak step down procedure will be used to determine thresholds for statistical significance for secondary analysis on the SEND pupil samples (Ludbrook, 1998).

Further secondary analysis is proposed for the full sample of pupils in the Year 9 and Year 8 cohorts. This sample will include all children in these year groups, within participating schools, for whom data are available, regardless of whether they have a SEND designation or otherwise.

For the full sample of students, the table below sets out the regression models that will form the basis of the secondary analysis for each Year group separately

Dependent variable	Model	Intervention group indicator	Region indicator	Further covariates	interference
GCSE English language Mark (standardised)	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) KS2 Reading raw score at pupil and school levels 2) Month of birth 3) Gender 4) FSM	Robust standard errors/95 per cent confidence interval
GCSE Mathematics Mark (standardised)	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) KS2 mathematics raw score at pupil and school levels 2) Month of birth 3) Gender 4) FSM	Robust standard errors/95 per cent confidence interval
Total difficulties (SDQ)	Hierarchical linear model - random effects at school and pupil levels	Yes	Yes	1) Total difficulties baseline score 2) Month of birth 3) Sex 4) FSM	Robust standard errors/95 per cent confidence interval

Results from these regression models will be reported as regression coefficient estimates in all cases and as effect sizes consistent with Hedge's g (Durlak, 2009).

Subgroup analyses

Subgroup analysis will examine the effect of AITT on English language GCSE standardised marks for those pupils ever-FSM. First a regression model of the following form will be estimated on each full year group sample using the `mixed` command in STATA v17 and maximum likelihood:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2(X_{ij} - \bar{X}_j) + \beta_3(\bar{X}_j - \bar{X}) + \beta_4 FSM_{ij} + \beta_5 FSM_{ij} \times T_j + \beta_6 R_j + \vartheta_j + \varepsilon_{ij}$$

Interest will centre on the sample estimate of β_5. Of interest is whether the 95 per cent confidence interval derived on the basis of robust standard errors leads the conclusion that the data are inconsistent with a value of zero for this parameter. A separate model for the ever-FSM subgroup only will also be estimated in a form identical to equation [1]. From this model we will report an effect size and 95 per cent confidence interval for the FSM subgroup consistent with Hedge g (see section below for derivation of effect sizes and details of this calculation). This separate model allows for the relationship between all covariates and the response to vary for the ever-FSM subgroup.

Longitudinal follow-up analyses

As discussed, two sets of analyses are proposed that will take be conducted roughly one year apart and reported separately. The first analyses will follow the approach set out above for pupils entering Year 9 at September 2020, both those designated SEND and all pupils in that year group cohort.

If the EEF decide that implementation of the intervention has been successful, it is proposed that the approach to primary and secondary analyses discussed above will be repeated using data from pupils entering Year 8 at September 2020. Analyses of these data might be considered medium to longer-term analyses.

Imbalance at baseline

We proposed to compare the characteristics of intervention and control group schools and pupils as measured at or prior to randomisation, in the 'as randomised' and 'as analysed' samples. The 'as randomised' sample will contain all schools and pupils at the point of randomisation that have not subsequently withdrawn from the study. The 'as analysed' sample for the Year 9 cohort will be all those pupils for whom we observe a GCSE English language mark at summer 2023 and values for the covariates implied by model [1] above. Likewise, the 'as analysed' sample for the Year 8 cohort will be all those pupils for whom we observe a GCSE English language mark at summer 2024 and values for the covariates implied by model.

We will present tabulations that compare the counts, proportions, means and standard deviations, as appropriate, for the 'as randomised' and 'as analysed' samples containing the following variables measured at, or prior to randomisation, for: the full sample, intervention and control group samples:

At the pupil level

- Gender
- Month of birth
- SEND status
- Ever free school meals
- KS2 points score English reading
- KS2 points score Mathematics
- Unauthorised absences 2019/20
- Authorised absences 2019/20
- Exclusions 2019/20

At the school level

- Region
- School size 2018/19
- Proportion of all students EAL 2018/19
- Proportion of all students ever-FSM 2018/19
- Proportion of all students SEND 2018/19
- Attainment 8 average score 2018/19,

These variables are chosen because they are used as covariates in the primary or secondary analysis discussed previously, with the exception of some of the school level variables. These are included primarily to examine the possible consequences of school drop out on the 'as randomised' sample.

Differences between intervention and control groups will be reported as standardised mean differences.

Missing data

Sensitivity tests examining the possible consequences of any missing data will be conducted for the primary analysis for both Year 8 and 9 reporting. The focus of this analysis will be an assessment of the extent of missingness that might frustrate intention to treat analysis on the basis of model [1] above, and therefore biased and/or imprecise sample estimates of β_1.

Missingness that occurs prior to randomisation is unlikely to cause bias in estimated treatment effects but can lead to diminished sample sizes. For the primary analysis the potential sources of missingness *prior* to randomisation would be:

1) Parents of SEND pupils removing their child from the study, in cases where the school as a whole continues to participate in the trial
2) Schools not supplying information that enable us to identify SEND pupils but which have provided other information about the pupil such that they are enumerated and considered a member of the 'as randomised' sample

3) Schools not supplying KS2 reading test score for pupils designated SEND but which have provided other information about the relevant pupils such that the pupil is enumerated and considered a member of the ‘as randomised’ sample

Missingness that occurs subsequent to randomisation can lead to diminished sample sizes, a loss of power and bias estimates of β_1. For the primary analysis potential sources of missingness subsequent to randomisation are likely to include:

1) Parents of SEND pupils removing their children from the study subsequent to randomisation and requesting the child’s data be deleted
2) Pupils leaving the school and moving to another setting outside of the trial sample – the study is not resourced to trace these children but we do not anticipate they will be large in number and there is no expectation that this will occur disproportionately by trial arm
3) Schools withdrawing from the study and requesting all data supplied by them be deleted
4) Schools failing to supply GCSE English marks and/or information about exam boards for pupils considered part of the ‘as randomised’ sample

The table below explores balance in the ‘as randomised’ sample cohorts for Years 8 and 9. These samples are subject to the loss of data prior to randomisation and the loss of one school from the ‘as randomised’ sample subsequent to randomisation. The situation can change in the future if more schools and pupils remove themselves from the study and ask for all their data to be destroyed. However, the present situation indicates that for the primary analysis missing data to this point do not appear to be particularly problematic. The differences between intervention and control groups in terms of missingness on the variable that indicates SEND status appears to be very small (for example 2.2 per cent in the intervention as compared to 1.4 per cent in the control arm for the Year 9 cohort). There are no missing values on the region indicator variable.

The percentage of cases in the ‘as randomised’ sample with missing values on their KS2 reading test score in both trial arms and for both year groups does exceed 5 per cent. However, the mean observed KS2 English reading test scores in the two groups are identical. There are also trivial amounts of missing data in the FSM indicator, a crucial variable for subgroup analysis.

What these provisional analyses reveal is that unless further schools and pupils remove themselves from the study and request that their data be destroyed, including observations on variables collected at the baseline, missing data prior to randomisation is unlikely to lead to sample sizes that are diminished to a troubling extent and the drop out of a single school to date does not appear to give rise to concerns. As a result, the challenge facing this study is more likely to come from missing data at follow-up; in this case missing observations on GCSE English language marks.

	Year 9			**Year 8**		
	Intervention	***Control***	***Diff***	***Intervention***	***Control***	***Diff***
Schools	78	78		78	78	
Pupils	14,934	14,768		14,814	15,455	
SEND status						
Observed	14,598	14,656		14,451	15,317	
Missing	336	112		363	138	
% Missing	2.2	0.8	1.4	2.5	0.9	1.6
SEND Pupils	1,924	2,257		2,359	2,514	
% of observed	13.2	15.4	-2.2	16.3	16.4	-0.1
SEND pupils by region (column percentage)						
East Midlands	12.0	12.0	0.0	12.3	11.0	1.3
East of England	3.5	8.5	-5.0	3.2	6.7	-3.5
London	9.7	9.1	0.6	10.8	9.9	0.9

North East	8.3	11.1	-2.8	9.9	9.6	0.3
North West	18.7	16.9	1.8	18.2	17.0	1.2
South East	7.2	4.3	2.9	5.3	5.5	-0.2
South West	23.9	16.6	7.3	19.5	17.8	1.7
West Midlands	14.1	15.7	-1.6	15.4	15.6	-0.2
Yorkshire & Humber	3.0	5.8	-2.8	5.7	7.2	-1.5
Total (N=)	1,924	2,257		2,359	2,514	
KS2 reading score						
Observed	14,157	13,917		13,929	14,502	
Missing	777	851		885	953	
% missing	5.2	5.8		6.0	6.2	
Ever-FSM status						
Observed	14,779	14,669		14,624	15,316	
Missing	155	99		190	139	
% missing	1.0	0.7		1.3	0.9	
Means:						
KS2 reading score (standard deviation)	33.5 (9.0)	33.4 (9.0)	0.08	33.3 (10.0)	33.3 (9.8)	-0.06

At analysis, in deciding whether missingness in the primary outcome – standardised GCSE marks in English language – is likely to be an important issue, we will first calculate the rate of missingness in the primary outcome, in both trial arms and compare them. This calculation will be based on the subset of the 'as randomised' sample for which a drop out model is estimated (see further below n = 55,520 at the time of writing). If the absolute level of missingness exceeds five per cent in both arms of the trial and the difference in this rate across arms exceeds 0.10 of a standardised difference then we propose to conduct full sensitivity analysis for missing data for the primary analysis.

The first step will be to model the determinants of missingness using available baseline information, the so called 'drop-out' model. For all sample members that supply the necessary data at baseline we propose to fit a multi-level logistic regression model where the response is a binary indicator revealing whether the GCSE mark is observed for a given pupil. The model will contain a random effect at the school level.

The following variables will be considered for inclusion as covariates in the drop-out model based on our prior expectations about which individual level factors might be associated with the failure to supply an observation on the primary outcome:

- Gender
- Month of birth
- Whether SEND
- Whether ever-FSM
- KS2 Reading test score
- Whether the child recorded any absences in 2018/19
- Whether the child had been excluded in 2018/19

At the school level and in a similar manner the following variables will be consider for inclusion in the model:

- School region
- School size in terms of number of pupils on the roll in 2018/19
- Proportion of the school roll that were SEND in 2018/19
- Proportion of the school roll that were ever-FSM in 2018/19

- Proportion of the school roll that were EAL in 2018/19

On the basis of these covariates, at the time of writing, taking into account missingness prior to randomisation and the drop out of one school subsequent to randomisation, the sample upon which the drop out model will be estimated comprises 27,400 pupils of the 30,223 in schools randomised to control (or 90% of the control group in the 'as randomised' sample) and 26,475 pupils of the 29,748 in schools randomised to the intervention (90% of the intervention group in the 'as randomised' sample). This is the sample available to us, at the time of writing, for modelling non-response and performing multiple imputation.

It is important to re-iterate, that although we propose performing multiple imputation (see below) for the primary outcome variable on a subset of the 'as randomised' sample, the missingness in this sample, if no further schools and/or pupils leave the trial and request their data to be destroyed, will be uncorrelated with the intervention group dummy variable in equation [1] T_i. This is because most of the missingness occurred prior to randomisation (with the exception of one school). This means conducting our missing data sensitivity analysis using this reduced sample will not introduce further bias with respect to estimating treatment effects. This conclusion is re-enforced by the analysis in the table immediately above; which shows that at randomisation our sample with respect to the response and covariates in the primary analysis is well balanced.

The drop out model will provide estimates of the association between these variables and the probability that an observation on GCSE English examination mark is missing. Covariates for which a parameter estimate is obtained with an associated 95 confidence interval that leads us to reject zero as population value for the parameter will be considered potential determinates of missingness in the primary outcome.

Subsequent to estimation of the drop-out model, we propose to sensitivity test the consequences of missing data in the primary outcome, on the assumption that they are missing at random (MAR), for the sample estimates of AITT in the primary analysis. An imputation model including the covariates discussed above and based on the `mice` package approach in `R` will be used to impute missing values on the primary outcome; mice is one of the few multiple imputation programs that can take account of clustered data. The burn-in phase for the imputation will consist of 20 imputation cycles with the number of imputed data sets set equal to the FMI (fraction of missing information) determined from an initial run, and with 20 cycles between the creation of each imputed data set. Imputation will be conducted in intervention and control groups separately. The stability of the imputation will be assessed through inspecting standard plots and tests. If these appear satisfactory, equation [1] will be estimated on the final merged imputed datasets as appropriate and results compared to the main primary analysis.

If multiple imputation does not appear to perform satisfactorily, then variables that appear to be associated with missingness at follow-up on the basis of results from fitting the drop out model will be added to the regression model [1] as additional covariates, and the model re-estimated on the completed cases at analysis sample file.

Compliance

Estimates of the average effect of intention to treat provide a perfectly valid and informative estimate of the effectiveness of an intervention. In some cases, however, interest is in the average effect of the intervention on those that participate or comply with their treatment assignment. This is particularly the case where there is interest in conducting an economic analysis. In such cases different forms of complier average causal effects (CACEs) become the estimands of interest.

WSS SEND review is a 'whole-school' intervention. This means that compliance is defined at the school level. If a school adheres to the trial protocol then all pupils in the school are 'treated' and are 'compliers'. For purpose of CACEs analysis we define a compling school as one that had engaged in the WSS Review initial training event. It is likely that some schools assigned to the intervention group will be non-compliant. Conversely, it is not possible for schools allocated to the control group to be non-compliant and to participate in WSS Review. This means that we face a situation of possible one-sided non-compliance (Gerber, Alan & Green, Donald, 2012). This means that CACEs can be interpreted as the average effect of treatment on the treated.

We propose to use instrumental variables estimation and two stage least squares to obtain estimates of CACEs. This involves estimating two equations. The first is a compliance equation in which take-up of initial training in WSS Review is modelled as a dependent variable with the treatment group indicator as a covariate. The fitted value for the dependent variable is then entered into a second stage equation, which is effectively model [1] above with the fitted values from the first stage equation replacing T_j. These two models can be modelled in a single step using the command '`ivregress 2sls`' in STATA v17, with standard errors adjusted for clustering of pupils within schools using the subcommand `vce(cluster robust).` Estimation of CACEs relies on the 'exclusion restriction' applying. This means that randomisation causes exogenous variation in compliance and the dependent variable, and it does so free from any confounding effects of third or unmeasured variables influencing both compliance and the outcome. In this case, the causal effects that are recovered are those on compliers only and by extention on those who take-up WSS Review.

Intra-cluster correlations (ICCs)

As described previous ICCs will be reported for all regression models estimated as part of the primary and secondary analysis. For the primary analysis, this includes a null or empty model that will yield an estimate of the full unconditional ICC for the primary outcome. This estimate will be obtained using the command `estat icc` in STATA v17

Effect size calculation

The primary outcome will already be standardised so that marks from different awarding bodies can be combined in to a single response variable. Standardisation of the primary outcome will mean that the standard deviations of the response in both intervention and control groups will be very close to one as will their variances.

The following equation represents the approach to deriving Hedges' g from a regression model (Hedges, 2007):

$$ES = J \times \frac{\hat{\beta}_1}{S} \times \sqrt{1 - \frac{2(n-1)\rho}{N-2}}$$

Here $\hat{\beta}_1$ is the AITT estimate from the regression model [1] above. Hedges (2007) shows that when cluster sizes are unequal, as they are here, substituting the average cluster size into the above is a close approximation to a much more complex equation for the effect size and its variance. In our case the average cluster size will be about 27 pupils and we assume rho is 0.2. The total sample size N will be approximately 4000. Thus we can calculate the term under the square root sign, and it is again very close to one and can be effectively be ignored, as can the factor J , a bias correction adjustment, often included in effect size calculations deriving Hedges g from Cohen's d which only applies when sample sizes are very small. S is the pooled within group standard deviation. Confidence intervals for the effect size will be derived on the basis of dividing through the 95 per cent upper and lower limits for the confidence intervals obtained from sample estimates of β_1 from equation [1] by S.

Bibliography

Dong, N., & Maynard, R. A. (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. https://doi.org/10.1080/19345747.2012.673143

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928.

Education Endowment Foundation. (2013). *Pre-testing in EEF evaluations*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/Pre-testing_paper.pdf

Gerber, Alan, S., & Green, Donald, P. (2012). *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, *25*(12), 1032–1037.

Morris, S., Lewin, C., Hick, P., Smith, A., & Harrison, J. (2020). *Evaluation of Whole School SEND Review: A cluster randomised controlled trial Evaluation Protocol*. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Whole_School_SEND_Protocol_Final.pdf

Morris, S. P., Smith, A., & Kiss, Z. (2018). *Statistical Analysis Plan Evaluating the effectiveness of Eedi formative assessment programme (previously Diagnostic Questions) on raising attainment in mathematics at GCSE*. https://educationendowmentfoundation.org.uk/public/files/Projects/EEDI_SAP_2018.11.29_FINAL.pdf

Morris, S., Smith, A., Lewin, C., Hick, P., & Harrison, J. (2021). Whole School SEND (WSS) Review: study protocol for a two-arm pragmatic parallel cluster randomised controlled trial in 160 English secondary schools. *Trials*, *22*(1), 333. https://doi.org/10.1186/s13063-021-05280-y