Rehearsal Room Writing – Efficacy Trial Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research (NFER)
Principal investigator(s): Helen Poet

PROJECT TITLE	**Rehearsal Room Writing – Efficacy Trial**
DEVELOPER (INSTITUTION)	The Royal Shakespeare Company (RSC)
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATOR(S)	Helen Poet
SAP AUTHOR(S)	Ruth Staunton, Helen Poet
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Year 5, 9-10 years old
NUMBER OF SCHOOLS	183 at design, 181 at randomisation
NUMBER OF TEACHERS	200 at design, 197 at randomisation
NUMBER OF PUPILS	5600 at design, 5113 at randomisation
PRIMARY OUTCOME MEASURE AND SOURCE	Writing Assessment Measure Comparative Judgement; Pupil responses to a writing prompt
SECONDARY OUTCOME MEASURE AND SOURCE	Lexical diversity, lexical sophistication, dependency grammar framework; Pupil responses to a writing prompt. Liking Writing Scale (LSW); Liking Writing survey. Self-efficacy for Writing Scale (SEWS); SEWS survey. Teacher Efficacy Scale for Writing; teacher efficacy for improving students' writing performance survey. KS2 writing, KS2 reading, KS2 maths; NPD.

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]		*N/A*

Table of contents

Introduction

Rehearsal Room Writing (RRW) is a programme that aims to improve Year 5 pupils' writing ability by increasing their skills, motivation, self-efficacy and enjoyment of creative writing through drama-based approaches to engaging with Shakespeare's texts. The programme was developed by the Royal Shakespeare Company (RSC) as part of their training offer for primary school teachers. Feedback from work with 280 state schools and 15 regional theatres across England led the RSC to consider the programme's potential for impact in relation to writing outcomes. While the RRW programme is not currently commercially available to schools, the RSC offers professional development and resources for teachers using some of the Rehearsal Room techniques that are implemented as part of the programme.

This trial aims to evaluate the efficacy of the RRW programme in improving writing ability, writing self-efficacy, enjoyment of writing and (longer term) Key Stage 2 assessment scores. It is a two-armed, school randomised trial. Schools will nominate one or two teachers to participate in the evaluation and nominated teachers in schools randomised to the intervention arm will attend 5 days of RRW CPD training. Teachers who have attended the training will deliver RRW sessions to their Y5 class(es) during the 2025/26 academic year. For the purpose of the trial, each teacher should deliver a minimum of 20 hours of lesson time to their Year 5 class using RRW approaches between November 2025 and June 2026. Pupil writing ability, writing self-efficacy and enjoyment of writing and teacher efficacy for improving students' writing performance were all measured at baseline (Autumn 2025) and will be measured again at endpoint (Summer 2026).

Design overview

Trial design, including number of arms		**Two-arm, cluster randomised**
Unit of randomisation		School
Stratification variables (if applicable)		None
Primary outcome	variable	Writing Ability
	measure (instrument, scale, source)	Writing Assessment Measure, Comparative Judgement true scores, pupil scripts
Secondary outcome(s)	variable(s)	i) Lexical diversity ii) Lexical sophistication iii) Noun phrase complexity iv) Pupils' enjoyment of writing v) Pupils' writing self-efficacy vi) Teacher self-efficacy in relation to teaching writing vii) KS2 writing viii) KS2 reading ix) KS2 maths
	measure(s)	i) Moving average type token ratio, pupil scripts

	(instrument, scale, source)	ii) Percentage of lemmas falling outside the most frequent 2,000 and the most frequent 3,000 words in English, 0-100, pupil scripts iii) Mean number of words per noun phrase and frequency of relative clauses, pupil scripts iv) Liking Writing Scale, pupil survey v) Self-efficacy for Writing Scale, pupil survey vi) Teacher Efficacy Scale for Writing, teacher survey vii) KS2 writing teacher assessment, binary Y/N pupil working at or at greater depth than the expected standard, NPD viii) KS2 reading scaled score, 80-120, NPD ix) KS2 maths scaled score, 80-120, NPD
Baseline for primary outcome	**variable**	Writing Ability
	measure (instrument, scale, source)	Writing Assessment Measure, Comparative Judgement true scores, pupil scripts
Baseline for secondary outcome	**variable**	i) Lexical diversity ii) Lexical sophistication iii) Noun phrase complexity iv) Pupils' enjoyment of writing v) Pupils' writing self-efficacy vi) Teacher self-efficacy in relation to teaching writing vii) Writing ability (same as baseline for primary outcome), viii) KS1 age related expectations in reading ix) KS1 age related expectations in maths
	measure (instrument, scale, source)	i) Moving average type token ratio, pupil scripts ii) Percentage of lemmas falling outside the most frequent 2,000 and the most frequent 3,000 words in English, 0-100, pupil scripts iii) Mean number of words per noun phrase and frequency of relative clauses, pupil scripts iv) Liking Writing Scale, pupil survey v) Self-efficacy for Writing Scale, pupil survey vi) Teacher Efficacy Scale for Writing, teacher survey vii) Writing Assessment Measure, Comparative Judgement true scores, pupil scripts viii) KS1 age related expectations in reading, categorical, NPD ix) KS1 age related expectations in maths, categorical, NPD

Sample size calculations overview

Since we only have one primary research question, no adjustment for multiple comparisons has been made and the significance threshold has been maintained at 5%.

Table 1 -Sample size calculations for protocol and randomisation stages, for all pupils (Overall; RQ1) and pupils eligible for free school meals (FSM; RQ4). Includes scenarios assuming no attrition and with expected attrition

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
		No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.188	0.200	0.204	0.219	0.190	0.201	0.206	0.222
Pre-test/ post-test correlations	level 1 (pupil)	0.68		0.68		0.68		0.68	
Intracluster correlations (ICCs)	level 2 (school)	0.19		0.19		0.19		0.19	
Alpha		0.05		0.05		0.05		0.05	
Power		0.8		0.8		0.8		0.8	
One-sided or two-sided?		Two-sided		Two-sided		Two-sided		Two-sided	
Average cluster size		30.6	26.0	8.9	7.6	28.2	24.0	8.2	7.0
Number of schools	intervention	92	82	92	82	91	82	91	82
	control	91	82	91	82	90	81	90	81
	total	183[1]	164	183	164	181	163	181	163
Number of pupils	intervention	2815	2133	819	621	2550	1951	742	568
	control	2785	2133	811	621	2563	1960	746	570
	total	5600	4266	1630	1242	5113	3911	1488	1138

	School level	Pupil level
Expected attrition (%)	10%	15%

Sample size calculations were undertaken using the *PowerUpR* package in R statistical software[2]. We have updated the post randomisation MDES based on the actual number of schools and pupils randomised but have continued to use an estimate for the FSM percentage as this information will only become available through the NPD.

Although this is an evaluation randomised at school level, the upper limit on the sample size in this evaluation was 200 teachers because the total capacity of the teacher training days was 100

[1] The target number of schools was updated from 183 to 184 part way through recruitment due to lower than anticipated average number of pupils per teacher. This is described in the protocol text, but no update was made to the table, and this table matches the protocol sample size table.

[2] R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

teachers and we intended to allocate equally between trial arms. From a statistical perspective we would prefer one teacher to participate per school to maximise the power/minimise the MDES. However, feedback from teachers suggested that the ways of working in some schools would make them disinclined to participate in the evaluation if only one of their Year 5 teachers could undertake the training (if assigned to the intervention arm). To ensure that no schools were dissuaded from participating which could have introduced a threat to generalisability of conclusions, we allowed in the sample size calculations that 9% of schools could enrol two teachers in the evaluation.

For the primary measure, the ICC and pre-post correlation are taken from the Year 5 cohort in the Helping Handwriting Shine trial which used a similar measure (Stone *et al.*, 2022). The average number of pupils per school (at protocol, overall, unadjusted for attrition) is the average KS2 class size in 23/24 (28) taken from data published by the Department for Education (DfE) (Department for Education, 2024b) multiplied by the anticipated number of teachers per school (1.09).

For the FSM subset calculation (secondary research question 4), the primary analysis ICC and pre-post correlation are used. The average number of pupils per school is 29.1% of the number of pupils expected in the primary measure. This is the FSM percentage among all Year 5 pupils in 23/24, taken from published DfE data (Department for Education, 2024).

Since only a subset of pupils will be analysed for some of our secondary measures, we have also undertaken a power calculation for one secondary measure (RQ2; moving average type token ratio) to determine the size of the subset. The ICC was determined through analysis of similar data, supplied by the University of Exeter. Unfortunately, no data was available for pre-post correlation for this calculation. Existing literature demonstrates that correlations of around 0.7 are common in educational research (Singh et al., 2023). We have selected 0.5 for this calculation as a conservative but not unrealistically small estimate. The results of these calculations are shown in the table below. Compared to the primary analysis, the lower ICC offsets the lower sample size giving a very reasonable MDES for this secondary analysis.

Table 2 - Sample size calculations for protocol and randomisation stages for RQ2. Includes scenarios assuming no attrition and with expected attrition

		Protocol		Randomisation	
		No attrition	Expected attrition	No attrition	Expected attrition
Minimum Detectable Effect Size (MDES)		0.163	0.178	0.164	0.178
Pre-test/ post-test correlations	level 1 (pupil)	0.5		0.5	
Intracluster correlations (ICCs)	level 2 (school)	0.09		0.09	
Alpha		0.05		0.05	
Power		0.8		0.8	
One-sided or two-sided?		Two-sided		Two-sided	
Average cluster size		10.9	9.3	10.9	9.3
Number of schools	intervention	92	82	91	82

	control	91	82	90	81
	total	183	164	181	163
Number of pupils	intervention	1005	762	990	757
	control	995	762	980	750
	total	2000	1524	1970	1507

Randomisation

The randomisation was a school level randomisation with equal allocation between groups. No stratification was implemented so simple randomisation was applied. Randomisation occurred after recruitment had closed and after teachers had been nominated by schools.

Randomisation used R statistical software and a seed was set for reproducibility. Code will be included in the appendix of the final report.

Analysis

The main analyses will be intention-to-treat and will follow the October 2022 EEF Statistical Analysis Guidance[3]. It will not be possible to blind analysts to group allocation due to difference in data structure between groups i.e. session delivery only recorded for intervention pupils. All analyses will be done using R statistical software.

Research questions

Primary

RQ1: How effective is RRW at improving the writing ability of Year 5 pupils?

Secondary

RQ2a: How effective is RRW at increasing the richness of Year 5 pupils' written vocabulary?

RQ2b: How effective is RRW at increasing the sophistication of Year 5 pupils' written vocabulary?

RQ3a: How effective is RRW at increasing the complexity of Year 5 pupils' use of noun phrases in writing?

RQ3b: How effective is RRW at increasing Year 5 pupils' use of relative clauses to complexify noun phrases in writing?

RQ4: How effective is RRW at improving the writing ability of Year 5 pupils eligible for FSM?

RQ5: How effective is RRW at improving Year 5 pupils' enjoyment of writing?

RQ6: How effective is RRW at increasing Year 5 pupils' writing self-efficacy?

[3] https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design

RQ7: How effective is RRW at improving the self-efficacy of Year 5 teachers in relation to teaching writing?

RQ8a: To what extent is pupil enjoyment of writing a mediator of the primary outcome?

RQ8b: To what extent is pupil writing self-efficacy a mediator of the primary outcome?

RQ9: How does the impact of RRW on pupils' writing attainment vary by the number of sessions taught by the teacher?

Longitudinal follow-up RQs:

RQ10a: What is the impact of RRW on pupils' writing attainment one year after the end of programme delivery (end of Year 6)?

RQ10b: What is the impact of RRW on the writing attainment of pupils eligible for FSM one year after the end of programme delivery (end of Y6)?

RQ11a: What is the impact of RRW on pupils' reading attainment one year after the end of programme delivery (end of Year 6)?

RQ11b: What is the impact of RRW on the reading attainment of pupils eligible for FSM one year after the end of programme delivery (end of Year 6)?

RQ12: In RRW schools, is pupil performance in maths at the end of Year 6 (one year after the end of the programme) no worse than for pupils in the control group? (non-inferiority analysis)

More detail on the RQs can be found in the published protocol: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/rehearsal-room-writing-trial

Primary outcome analysis (RQ1)

Measures

The primary outcome measure will be a Writing Assessment Measure assessed using comparative judgement (WAM_CJ). We will work with No More Marking (NMM) to generate this measure. The measure aligns with the long-term outcome theorised by the logic model. Assessment of writing is multi-dimensional and contested (Clarkson, 2024), and the most commonly used form of assessment, rubric-based scoring, has been shown to be less accurate and consistent than the increasingly used Comparative Judgement (Pollitt, 2012; Pinot de Moria, Wheadon and Christodoulou, 2022). Pupils will be presented with a narrative writing pictorial prompt at both baseline and endpoint (a picture from the NMM archive, unrelated to Shakespeare). They will be asked to write for up to 30 minutes. Scripts (from both baseline and endpoint) will be scanned into the NMM platform and external, allocation-blinded judges (recruited and managed by NFER) will evaluate the scripts. The NMM process uses a Bradley-Terry model (Hunter, 2004, pp. 384-406) to generate true scores from the win/loss data provided by judges when pairs of scripts are presented to them. True scores in our study will be a measure of latent writing ability. They are expected to follow a Normal distribution and will be standardised to have a mean of 0 and standard deviation of 2. Pupils will be included in the primary analysis if

they attend a randomised school at endpoint and complete the endpoint writing task. Pupils missing a baseline score will be imputed at the mean (expected to be 0 due to standardisation).

Analysis

The primary outcome measure of WAM_CJ will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model for RQ1 can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

Heterogeneity of variance will be assessed for all models through visual inspection of the distribution of residuals. Should the assumption fail to be upheld, transformation of the outcome will be investigated, or a non-parametric alternative model will be applied.

Secondary outcome analysis (RQ2, RQ3, RQ5, RQ6, RQ7)

Measures

The outcome measures for RQ2 and RQ3 will be evaluated from the same scripts used to generate the primary outcome measure.

A random subset of 10 scripts per class will be selected for transcription to allow analysis of the RQ2 measures. Sampling for RQ2 will be stratified by FSM eligibility i.e. it will be performed separately within FSM eligible pupils and pupils not known to be eligible for FSM, with a proportionate number of pupils selected in each category to ensure that the percentage of FSM pupils is accurately represented in the subset of pupils. Due to timeline, the FSM variable used for this stratification will be collected from schools and will be current FSM eligibility, in contrast to the FSM variable used for the subgroup analysis (RQ4) which will be sourced from the NPD and will be 'recorded as eligible for FSM in the last 6 years'.

One script per class will be randomly selected for generation of the RQ3 measures from within the RQ2 subset.

Richness of vocabulary use refers to two constructs: lexical diversity and lexical sophistication. Lexical diversity (**RQ2a**) refers to the extent to which writers used a range of different words, rather than repeating items. This is known to correlate with both writer age and teachers' evaluation of writing quality (Durrant, Brenchley and McCallum, 2021). Lexical diversity will be measured using the moving average type-token ratio (MATTR). This is found by calculating Type Token Ratios (TTR - i.e. the number of distinct words divided by the total number of words) for successive N-word segments of the text. For example, if using 50-word segments, TTRs would be calculated for words 1-50, 2-51, 3-52, etc. Until the end of the text is reached). MATTR is the arithmetic mean of these TTRs. Research has shown that this provides a robust way of quantifying the number of different words that writers used independently of text length (Zenker and Kyle, 2021), so enabling fair comparison of texts of different lengths. For this analysis we have opted to use 50 word segments so our outcome measure is MATTR50.

Lexical sophistication (**RQ2b**) refers to the extent to which writers use vocabulary that is not frequent in the language as a whole. This also correlates with both writer age and teachers' evaluation of writing quality (Durrant, Brenchley and McCallum, 2021). Lexical sophistication is often operationalized in terms of the percentage of all words in a text that do not appear on a list of high-frequency vocabulary. For greater sensitivity, two such measures will be used in this project: the percentage of lemmas[4] falling outside the most frequent 2,000 and the most frequent 3,000 words in English, as evidenced by the most recent edition of the British National Corpus (Brezina, Hawtin and McEnery, 2021).

Noun phrase complexity (**RQ3**) refers to the average number of components included within each noun phrase. This is operationalized as the mean number of words per noun phrase (**RQ3a**), a measure which has been shown to correlate with learner age (Durrant and Brenchley, 2023). Previous research has highlighted the use of relative clauses as a key site of complexity development within noun phrases (Durrant and Brenchley, 2023), so the frequency of this form is studied separately as a potential area of development (**RQ3b**). Noun phrases and their internal components will be initially identified by an automated parser (Manning *et al.*, 2014). The output of the parser will then be manually checked for accuracy by analysts who will be trained by the project team.

The pupil surveys at baseline and endpoint will include items from the Self-Efficacy for Writing Scale and items from the Liking Writing Scale (Bruning et al., 2013). These are existing validated scales, suitable for this age group. Questions will be read aloud to pupils by the teacher. Due to small wording changes to the questions which were considered essential for appropriate use in a UK classroom, our initial plan to apply the factor loadings in Bruning et al., 2013, was revised. Baseline responses to the items on each of these scales will undergo (separate) factor analyses and the factor from each which explains the most variance will be the baseline measure for RQ5 and RQ6. The loadings from these factor analyses will be applied to the endpoint survey responses to form the outcome measures for **RQ5** (pupils' enjoyment of writing) and **RQ6** (pupils' writing self-efficacy).

[4] *lemmas* combine all grammatically inflected forms of a word into a single unit; i.e. single and plural nouns (e.g., *dog-dogs; child-children*) and different forms of a verb (e.g. *like-likes-liked-liking*) are each treated as instances of the same lemma. However, they distinguish between words with the same spelling but different parts of speech (e.g. the noun *table* in *lay your cards on the table* is not the same lemma as the verb *table* in *she tabled a proposal*

The outcome measure for **RQ7** is teacher self-efficacy in relation to teaching writing. In the teacher surveys at baseline and at endpoint, teachers will be asked four items from the teacher efficacy for improving students' writing performance scale (Gilbert and Graham, 2010). Responses to the items on this scale at baseline will undergo a factor analysis and the factor which explains the most variance will be the baseline outcome for RQ7. The loadings from this factor analysis will be applied to the endpoint responses to form the outcome measures for RQ7.

All endpoint measures for **RQ2**, **RQ3**, **RQ5**, **RQ6** and **RQ7** were recorded at baseline before randomisation via a baseline writing assessment (also from NMM; this used a different pictorial prompt to the one that will be used at endpoint) and baseline pupil and teacher surveys.

Analysis

RQ2a: The secondary outcome measure of moving average type token ratio will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores, FSM (stratification variable used in selection of the sub-sample for this analysis) and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be a randomly selected subset of 10 pupils per class. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 FSM_{ij} + \beta_4 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint MATTR50 for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the coefficient for the FSM variable used as a stratifier in the selection of the pupil subset.
- β_4 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline MATTR50 for pupil *i* in school *j*.
- FSM_{ij} is the current eligibility for FSM (the stratification variable used in the selection of the pupil subset).
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ2b: The secondary outcome measures of percentage of lemmas falling outside the most frequent 2,000 and the most frequent 3,000 words in English will be used as dependent variables in two linear mixed effects model with intervention group as a predictor, controlling for baseline scores, FSM (stratification variable used in selection of the sub-sample for this analysis) and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be a randomly selected subset of 10 pupils per class (the same subset used for RQ2a). The models can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 FSM_{ij} + \beta_4 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint percentage of lemmas which fall outside the most frequent 2000 or 3000 words in English for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the coefficient for the FSM variable used as a stratifier in the selection of the pupil subset.
- β_4 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline percentage of lemmas falling outside the most frequent 2000 or 3000 words in English for pupil *i* in school *j*.
- FSM_{ij} is the current eligibility for FSM (the stratification variable used in the selection of the pupil subset).
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ3a: The secondary outcome measure of mean number of words per noun phrase will be used as the dependent variable in a linear model with intervention group as a predictor, controlling for baseline scores. The analysis population will be one randomly selected pupil per class. The model can be represented as:

$$Y_i = \beta_0 + \beta_1 InterventionGroup_i + \beta_2 BaselineScore_i + \epsilon_i$$

Where:

- Y_i is the endpoint mean number of words per noun phrase for pupil *i*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- $InterventionGroup_i$ is the intervention group that the school attended by pupil *i* was randomly assigned to.
- $BaselineScore_i$ is the baseline mean number of words per noun phrase for pupil *i*.
- ϵ_i is the residual error for pupil *i*.

RQ3b: The secondary outcome measure of frequency of relative clauses will be used as the dependent variable in a linear model with intervention group as a predictor, controlling for baseline scores. The analysis population will be one randomly selected pupil per class (the same pupil as used in RQ3a). The model can be represented as:

$$Y_i = \beta_0 + \beta_1 InterventionGroup_i + \beta_2 BaselineScore_i + \epsilon_i$$

Where:

- Y_i is the endpoint frequency of relative clauses for pupil *i*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- $InterventionGroup_i$ is the intervention group that the school attended by pupil *i* was randomly assigned to.
- $BaselineScore_i$ is the baseline frequency of relative clauses for pupil *i*.
- ϵ_i is the residual error for pupil *i*.

[Note RQ4 is covered below in subgroup analysis]

RQ5: The secondary outcome measure of pupils' enjoyment of writing will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the pupils' enjoyment of writing at endpoint for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the pupils' enjoyment of writing at baseline for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ6: The secondary outcome measure of pupils' writing self-efficacy will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the pupils' writing self-efficacy at endpoint for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.

- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the pupils' writing self-efficacy at baseline for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ7: The secondary outcome measure of teacher self-efficacy in relation to teaching writing (average value per school) will be used as the dependent variable in a linear model with intervention group as a predictor, controlling for baseline scores. The analysis population will be all schools included in the evaluation. The model can be represented as:

$$Y_i = \beta_0 + \beta_1 InterventionGroup_i + \beta_2 BaselineScore_i + \epsilon_i$$

Where:

- Y_i is the teacher self-efficacy in relation to teaching writing at endpoint for school *i*. This is averaged across teachers within a school where two teachers are in the trial.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- $InterventionGroup_i$ is the intervention group that the school *i* was randomly assigned to.
- $BaselineScore_i$ is the teacher self-efficacy in relation to teaching writing at baseline for school *i*.
- ϵ_i is the residual error for school *i*.

Subgroup analysis (RQ4)

Measures

The subgroup analysis will use the same endpoint and baseline measures as RQ1. The subgroup definition will be a binary variable indicating if the pupil has been eligible for FSM in the preceding 6 years. This variable is available through the National Pupil Database (EVERFSM_6_P) so this analysis will be performed within the ONS Secure Research Service (SRS).

Analysis

RQ4: The primary outcome of WAM_CJ will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be pupils who have been recorded as eligible for FSM in the last 6 years. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

An additional model will be run repeating RQ1 but with the addition of an interaction term of intervention group by FSM eligibility. This model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 FSM_{ij} + \beta_3 InterventionGroup_j : FSM_{ij} + \beta_4 BaselineScore_{ij} + \beta_5 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the FSM variable.
- β_3 is the coefficient for the intervention group by FSM variable interaction.
- β_4 is the coefficient for the baseline score.
- β_5 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- FSM_{ij} is the binary variable indicating whether pupil *i* in school *j* has been eligible for FSM in the preceding 6 years.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

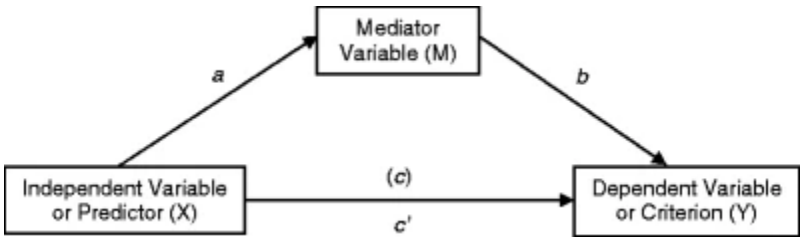
Additional analyses (RQ8, RQ9)

Measures

The analyses for RQ8 and RQ9 will use the same endpoint and baseline measures as RQ1. The dosage measure for RQ9 will be the number of RRW sessions delivered to each pupil.

Analysis

Mediation analyses are described using the following notation:



RQ8a: The secondary outcome of pupils' enjoyment of writing (M) will be explored as the mediator in a basic mediation model (Pieters, 2017), with the primary outcome of WAM_CJ as the dependent variable (Y) and the intervention group as the independent variable (X). The analysis population will be all pupils included in the evaluation.

The direct or c-path model can be represented as:

$$Y_{ij} = \beta_{0C} + \beta_{1C} InterventionGroup_j + b_{0Cj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_{0C} is the intercept.
- β_{1C} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- b_{0Cj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The regression of the mediator on the independent variable or a-path model can be represented as:

$$Y_{ij} = \beta_{0A} + \beta_{1A} InterventionGroup_j + b_{0Aj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the pupils' enjoyment of writing at endpoint for pupil *i* in school *j*.
- β_{0A} is the intercept.
- β_{1A} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- b_{0Aj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The mediated model can be represented as:

$$Y_{ij} = \beta_{0M} + \beta_{1M} InterventionGroup_j + \beta_{2M} EnjoyWriting_{ij} + b_{0Mj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_{0M} is the intercept.

- β_{1M} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_{2M} is the coefficient of the mediator, pupils' enjoyment of writing.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $EnjoyWriting_{ij}$ is the enjoyment of writing score for pupil *i* in school *j*.
- b_{0Mj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The key outcome for this analysis will be the total indirect effect of the mediator which can be calculated from these models as $\beta_{1A} * \beta_{2M}$ (the intervention group coefficient in the regression of the mediator on the intervention group multiplied by the mediator coefficient in the mediated model). We will report the magnitude of all model coefficients along with test statistics and *p*-values. We will also report the impact on the independent variable caused by including the mediator in the analysis i.e. $\beta_{1C} - \beta_{1M}$, (the intervention group coefficient in the direct model minus the intervention group coefficient in the mediated model). This will be reported as a sensitivity check as it should be similar in magnitude to the total indirect effect.

RQ8b: The secondary outcome of pupils' writing self-efficacy will be explored as the mediator in a basic mediation model, with the primary outcome of WAM_CJ as the dependent variable. The analysis population will be all pupils included in the evaluation.

The direct or c-path model is the same as that used in RQ8a and can be represented as:

$$Y_{ij} = \beta_{0C} + \beta_{1C} InterventionGroup_j + b_{0Cj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_{0C} is the intercept.
- β_{1C} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- b_{0Cj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The regression of the mediator on the independent variable or a-path model can be represented as:

$$Y_{ij} = \beta_{0A} + \beta_{1A} InterventionGroup_j + b_{0Aj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the pupils' writing self-efficacy at endpoint for pupil *i* in school *j*.
- β_{0A} is the intercept.
- β_{1A} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- b_{0Aj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The mediated model can be represented as:

$$Y_{ij} = \beta_{0M} + \beta_{1M} InterventionGroup_j + \beta_{2M} WritingSelfEfficacy_{ij} + b_{0Mj} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_{0M} is the intercept.
- β_{1M} is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_{2M} is the coefficient of the mediator, pupils' writing self-efficacy.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $WritingSelfEfficacy_{ij}$ is the writing self-efficacy score for pupil *i* in school *j*.
- b_{0Mj} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

The key outcome for this analysis will be the total indirect effect of the mediator which can be calculated from these models as $\beta_{1A} * \beta_{2M}$, (the intervention group coefficient in the regression of the mediator on the intervention group model multiplied by the mediator coefficient in the mediated model). We will report the magnitude of all model coefficients along with test statistics and *p*-values. We will also report the impact on the independent variable caused by including the mediator in the analysis i.e. $\beta_{1C} - \beta_{1M}$, (the intervention group coefficient in the direct model minus the intervention group coefficient in the mediated model). This will be reported as a sensitivity check as it should be similar in magnitude to the total indirect effect.

RQ9: The primary outcome of WAM_CJ will be used as the dependent variable in a linear mixed effects model with number of sessions as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all intervention group pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 Dosage_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the endpoint WAM_CJ for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest representing the difference in Y for an increase of one RRW session.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $Dosage_j$ is the number of RRW sessions delivered to pupil *i* in school *j*.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

Longitudinal follow-up analyses (RQ10, RQ11, RQ12)

Measures

The outcomes for RQ10a, RQ10b, RQ11a, RQ11b and RQ12 are all KS2 outcomes held in the National Pupil Database so these analyses will be performed within the ONS SRS. The study population is Year 5 pupils in the 2025/26 academic year, who will take the end of KS2 assessments in summer 2027. KS2 outcomes will be available for all of these pupils who were present on the day of the exam in the autumn of 2027. These longitudinal outcomes will be analysed and reported as an addendum to the original report.

For RQ10a and RQ10b, the outcome for the writing teacher assessment (KS2_WRITTAOUTCOME) will be used to create a binary outcome measure where 1 indicates that the pupil is either 'working at the expected standard' or 'working at greater depth within the expected standard' (i.e. 1 indicates KS2_WRITTAOUTCOME is 'EXS' or 'GDS', 0 indicates any other code recorded for this variable).

For RQ11a and RQ11b, the scaled score in reading (KS2_READSCORE) will be used as the outcome measure.

For RQ12, the scaled score in maths (KS2_MATSCORE) will be used as the outcome measure.

For baseline measures, RQ10a, RQ10b, RQ11a and RQ11b will use the WAM_CJ baseline score from NMM, as in RQ1. RQ12 will use Year 4 Multiplication Tables Check score (MTC_FormMark) as the baseline measure. As an exploratory analysis, additional models for RQ11a and RQ11b will be run using categorical KS1 results, indicating attainment against age related expectations in reading (KS1_READ_OUTCOME) as the baseline measure for each pupil, and two additional models for RQ12 will be run using categorical KS1 results, indicating attainment against age related expectations in maths (KS1_MATH_OUTCOME) and WAM_CJ baseline score as the baseline measure for each pupil. From these models we will report the variance explained by baseline score and compare this to the variance explained by baseline score in the models that use WAM_CJ as baseline. We note that the pupils in our study will have been the last cohort to take KS1 statutory assessments when they were compulsory (May 2023).

Analysis

RQ10a: The secondary outcome measure of KS2 writing outcome will be used as the dependent variable in a binomial generalised linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model can be represented as:

$$ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j}$$

Where:

- p_{ij} is the probability that pupil *i* in school *j* was working at the expected standard or working at greater depth within the expected standard in the KS2 writing teacher assessment.

- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in log odds between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.

RQ10b: The secondary outcome measure of KS2 writing outcome will be used as the dependent variable in a binomial generalised linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be pupils who have been recorded as eligible for FSM in the last 6 years. The model representation is identical to the one specified above for RQ10a.

RQ11a: The secondary outcome measure of KS2 reading scaled score will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the KS2 reading scaled score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ11b: The secondary outcome measure of KS2 reading scaled score will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor, controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be pupils who have been recorded as eligible for FSM in the last 6 years. The model representation is identical to the one specified above for RQ11a.

RQ12: The secondary outcome measure of KS2 maths scaled score will be used as the dependent variable in a linear mixed effects model with intervention group as a predictor,

controlling for baseline scores and class as covariates, and accounting for clustering of pupils within schools. The analysis population will be all pupils included in the evaluation. The model can be represented as:

$$Y_{ij} = \beta_0 + \beta_1 InterventionGroup_j + \beta_2 BaselineScore_{ij} + \beta_3 Class_{ij} + b_{0j} + \epsilon_{ij}$$

Where:

- Y_{ij} is the KS2 maths scaled score for pupil *i* in school *j*.
- β_0 is the intercept.
- β_1 is the coefficient of interest for the intervention group representing the difference in Y between the two groups.
- β_2 is the coefficient for the baseline score.
- β_3 is the vector of class coefficients.
- $InterventionGroup_j$ is the intervention group that school *j* was randomly assigned to.
- $BaselineScore_{ij}$ is the baseline WAM_CJ for pupil *i* in school *j*.
- $Class_{ij}$ is the class that pupil *i* in school *j* is taught in.
- b_{0j} is the random intercept for school *j*.
- ϵ_{ij} is the residual error for pupil *i* in school *j*.

RQ12 will be tested against a non-inferiority limit. See Effect size section for more detail.

Imbalance at baseline

Tables showing the balance between intervention groups for all randomised schools and pupils and for the primary analysis population will be reported. These tables will present the number and percentage of pupils or schools for the following (categorical) characteristics: eligible for FSM, establishment type, geographical region (GOR), urban/rural and Ofsted rating. The tables will present means and standard deviations for the following (continuous) characteristics: pupil to qualified teacher ratio, KS1 to KS2 progress measures and baseline WAM_CJ score. The difference in the baseline score between the intervention and control will be reported as an effect size. The baseline scores will also be presented as histograms for the whole population and separately for the intervention and the control group.

Missing data

The number and proportion of pupils with missing endpoint WAM_CJ scores (primary outcome variable) will be reported. If the percentage of pupils present at randomisation but missing this outcome variable is less than 5%, no further missing data analysis will take place. If the percentage of pupils missing is greater than 5%, a logistic multilevel model will be run with a binary outcome variable indicating missing endpoint WAM_CJ. This model will include the intervention group variable as a predictor, along with the following pupil and school characteristics: gender, FSM eligibility, establishment type, school size (number of pupils), geographical region, urban/rural. Any of the additional variables which demonstrate a significant association with missingness will be included as a covariate in a rerun of the RQ1 analysis as a sensitivity check.

Compliance

We will use three definitions of compliance to create three school-level compliance metrics. Minimal compliance in training will be a binary measure defined as having attended at least three days of training, two of which must be the first two days. Optimal compliance in training will be a binary measure defined as having attended all five training days. Where more than one teacher was nominated by the school, both teachers must achieve the definition for the school to be considered compliant. Compliance in delivery will be a continuous measure defined as the number of hours delivered to pupils. For each of the compliance metrics, a complier average causal effect (CACE) analysis will be undertaken using a two stage least squares instrumental variable approach. These three analyses will be independent of each other. We recognise that the optimal school training compliance measure risks violating the CACE analysis exclusion restriction as we would expect at least some effect of teachers attending fewer than 5 training sessions. The appropriate caveats will be included in reporting and the primary focus for reporting the teacher training compliance will be on the minimal compliance measure.

For the first stage in each of the three compliance analyses, the compliance indicator will be regressed on the intervention group, together with the covariate from the primary analysis model (baseline WAM_CJ score). This first stage linear regression model will be:

$$compliance_i = \beta_0 + \beta_1 InterventionGroup_i + \beta_2 BaselineScore_i + \epsilon_i$$

Where:

- $compliance_i$ is the compliance measure for school *i*.
- β_0 is the intercept.
- β_1 is the intervention group coefficient.
- β_2 is the coefficient for the baseline score.
- $InterventionGroup_i$ is the intervention group that school *i* was randomly assigned to.
- $BaselineScore_i$ is the average baseline WAM_CJ score for school *i*.
- ϵ_i is the residual error

The compliance indicator is expected to take the value zero for all pupils in the control group (one-sided non-compliance).

For the second stage, endpoint WAM_CJ will be regressed on each pupil's predicted compliance value, $\widehat{compliance}_i$ (estimated from the first stage model), in the following linear regression model:

$$Y_i = \beta_0 + \beta_1 \widehat{compliance}_i + \beta_2 BaselineScore_i + \epsilon_i$$

Where:

- Y_i is the average endpoint WAM_CJ for school *i*.
- β_0 is the intercept.
- β_1 is the predicted compliance coefficient.
- β_2 is the coefficient for the baseline score.
- $\widehat{compliance}_i$ is the predicted compliance for school *i* from the first stage linear model.
- $BaselineScore_i$ is the average baseline WAM_CJ score for school *i*.

- ϵ_i is the residual error

The coefficient for predicted compliance β_1 in the second stage model is the CACE (complier average causal effect) estimate for the effect of compliance on endpoint WAM_CJ. Results from both regression stages will be reported for all three analyses.

Exploratory analysis

In addition, we are interested in the distribution of the timing and intensity of session delivery across schools in relation to the training sessions, and any impact this may have on the outcomes. We will report summary statistics such as the median start date, half-way through date and end date of delivery, and the mean & standard deviation of the number of sessions per week (between the within-school start and end dates). We will also run a compliance sensitivity analysis. This will repeat the minimal compliance training CACE analysis but with an additional criterion. For a school to be compliant teachers must

- have attended at least three days of training, two of which must be the first two days (minimal compliance metric as described above)
- have delivered at least 10 hours of the sessions after the teacher attended their third training day.

Intra-cluster correlations (ICCs)

The ICCs (school and residual) will be calculated as the proportion of the total model variance attributed to each level. These will be reported for the primary analysis model and for an unadjusted model with only the intervention group as a fixed effect.

Effect size calculation

Effect sizes will be calculated by dividing the adjusted difference in means (e.g. β_1 from the primary analysis model) by the square root of the total variance from the unadjusted model (with only intervention group as the fixed effect).

$$ES = \frac{\beta_1}{\sqrt{\sigma_T^2}}$$

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2$$

Where β_1 is the intervention group coefficient from the primary analysis model, σ_T^2 is the total variance from the unadjusted model, σ_B^2 is the between school variance from the unadjusted model and σ_W^2 is the within school variance from the unadjusted model.

Effect sizes will be significance tested against 0 using a 5% threshold for significance and will be reported with 95% confidence intervals and p-values. For RQ10, effects will be reported as odds ratios and tested against 1. For RQ12, the effect will be tested against a non-inferiority limit. This limit will be 10% of the outcome scale (scale is 80 to 120 so 10% is 4) divided by the square root of the total variance in the unadjusted model to bring the limit onto the effect size scale.

References

Brezina, V., Hawtin, A. and McEnery, T. (2021) 'The Written British National Corpus 2014 – design and comparability', Text & Talk, 41(5–6), pp. 595–615. Available at: https://doi.org/10.1515/text-2020-0052.

Bruning, R., Dempsey, M., Kauffman, D., McKim, C. and Zumbrunn, S. (2013) 'Examining dimensions of self-efficacy for writing', Journal of Educational Psychology, 105(1), pp. 25–38. Available at: https://doi.org/10.1037/a0029692.

Clarkson, R. (2024) '"It's missing the heart of what writing is about": teachers' interpretations of writing assessment criteria', British Educational Research Journal, 50(1), pp. 134–161. Available at: https://doi.org/10.1002/berj.3916.

Department for Education (2024) Schools, pupils and their characteristics, Academic year 2023/24, GOV.UK. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics (Accessed: 28 March 2025).

Durrant, P. and Brenchley, M. (2023) 'Development of Noun Phrase Complexity Across Genres in Children's Writing', Applied Linguistics, 44(2), pp. 239–264. Available at: https://doi.org/10.1093/applin/amac032.

Durrant, P., Brenchley, M. and McCallum, L. (2021) Understanding development and proficiency in writing: quantitative corpus linguistic approaches. Cambridge: Cambridge University Press.

Gilbert, J. and Graham, S. (2010) 'Teaching Writing to Elementary Students in Grades 4-6: A National SUrvey', The Elementary School Journal, 110(4).

Hunter, D.R. (2004). MM algorithms for generalized Bradley-Terry models. Annals of Statistics 32(1).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. (2014) 'The Stanford CoreNLP Natural Language Processing Toolkit', in K. Bontcheva and J. Zhu (eds) Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics, pp. 55–60. Available at: https://doi.org/10.3115/v1/P14-5010.

Pieters, R. (2017) 'Meaningful mediation analysis: plausible casual inference and informative communication', Journal of Consumer Research, 44(3), pp. 692–716.

Pinot de Moria, A., Wheadon, C. and Christodoulou, D. (2022) 'The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment', Research in Education, 113(1), pp. 25–40. Available at: https://doi.org/10.1177/00345237221118116.

Pollitt, A. (2012) 'The method of Adaptive Comparative Judgement', Assessment in Education: Principles, Policy & Practice, 19(3), pp. 281–300. Available at: https://doi.org/10.1080/0969594X.2012.665354.

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Higgins, S., Einbeck, J., Culliney, M. and Demack, J. (2023) Improving Power Calculations in Educational Trials. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Work_Package_2023-WP6_18_09_2023_FINAL.pdf?v=1713850351.

Stone, G., Andrade, J., Martin, K. and Styles, B. (2022) Helping Handwriting SHine - Evaluation Report. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Helping-Handwriting-Shine-Addendum-Report-Final.pdf?v=1743091937.

Zenker, F. and Kyle, K. (2021) 'Investigating minimum text lengths for lexical diversity indices', Assessing Writing, 47, p. 100505. Available at: https://doi.org/10.1016/j.asw.2020.100505.