Writing Roots
Statistical Analysis Plan

Evaluator (institution): RAND Europe, University of Leeds
Principal investigator(s): Elena Rosa Speciani

PROJECT TITLE	Using the Writing Roots programme to improve writing attainment: a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	Literacy Tree
EVALUATOR (INSTITUTION)	RAND Europe and University of Leeds
PRINCIPAL INVESTIGATOR(S)	Elena Rosa Speciani
PROTOCOL AUTHOR(S)	Merrilyn Groom, Anna Brian, Maria Jose Guevara, James Merewood
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	6 – 7 (Year 2, KS1), 9 – 10 (Year 5, KS2)
NUMBER OF SCHOOLS	110
NUMBER OF PUPILS	3657
PRIMARY OUTCOME MEASURE AND SOURCE	Writing attainment, Writing Assessment Measure (WAM)
SECONDARY OUTCOME MEASURE AND SOURCE	Writing self-efficacy, Self-Efficacy for Writing Scale

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	03/08/2026	*N/A*

Table of contents

Introduction

This trial is being conducted to explore how an adaption to the traditional writing curriculum in primary schools impacts pupils' writing attainment. Writing attainment has been a persistent concern in UK education, with recent Ofsted reports highlighting the need for innovative approaches to improve literacy outcomes (Ofsted, 2022). The Department for Education (DfE) has emphasised the importance of evidence-based interventions in addressing literacy gaps, particularly in the wake of disruptions caused by the COVID-19 pandemic (DfE, 2021).

Evidence suggests that targeted interventions can improve writing outcomes (for example, Graham & Perin, 2007; Graham et al., 2012; Finalayson & McCruddon, 2019), particularly when they are implemented systematically across schools (Graham et al., 2012). This trial examines an intervention that increases the diversity of writing conventions encountered by pupils and encourages pupils to engage with them creatively in their writing through a book-based pedagogical framework. Given evidence that that diversity in classroom texts enhances pupil engagement and motivation (O'Leary et al., 2024), and that motivation and enjoyment in writing are strongly correlated with improved writing outcomes, as noted by Ofsted (2022) and Slavin et al., (2019), there is potential for this approach to create impact. Despite this, there is little evidence on how this approach to teaching impacts writing attainment, underscoring the need for a rigorous evaluation.

Writing Roots is a whole-school writing programme developed by Literacy Tree. The programme takes a book-based approach to teaching writing, exposing pupils to a diverse range of high-quality children's texts and encouraging them to apply writing conventions creatively for different audiences and purposes. The programme aims to improve pupils' writing attainment, broaden vocabulary, and increase motivation and confidence in writing.

Teachers receive training through a half-day INSET launch event and three year-group-specific Teach Through a Text online training sessions across the academic year. Schools are also provided with access to the Literacy Tree platform, which offers detailed lesson plans, recommended texts, and planning support. Teachers deliver Writing Roots as daily one-hour writing lessons using structured sequences that integrate modelling, discussion, and opportunities for extended writing. All schools benefit from a link consultant, who is their primary contact for all pedagogical matters and implementation queries. Link consultants touch base with teachers in the first month, third month and ninth month, with ad hoc support provided as needed. Schools are offered the opportunity to visit Flagship Schools, which are schools selected by Literacy Tree to serve as high-fidelity examples of Writing Roots delivery, to allow teachers to see Writing Roots implemented in practice. Flagship Schools have been using Literacy Tree for at least 2 years and act as ambassadors for the approach.

The programme is widely used in England, with over 1,100 schools having accessed Writing Roots resources. However, it has not previously undergone an independent large-scale evaluation. This trial will assess the programme's impact using a two-arm cluster randomised controlled design, with randomisation at school level. Settings in the treatment group will receive the intervention throughout the 2025/2026 academic year, with control settings continuing with business-as-usual writing curriculum provision. Schools in the intervention group will implement the

programme across Years 1–6, whilst data collection and analysis is only carried out for Year 2 (Key Stage (KS) 1) and Year 5 (KS2). Baseline testing for most schools was conducted in the summer term of the 2024/2025 academic year, with schools in randomisation block 3 (see Randomisation section) being baseline tested in autumn 2025/2026. Endline testing for all schools will take place in the summer term of the 2025/2026 academic year. The evaluation will measure writing proficiency as the primary outcome and individual writing skills and self-efficacy as the secondary outcomes.

Design overview

Table 1: design overview

Trial design, including number of arms		Two-armed randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		Education Investment Area
Primary outcome	Variable	Writing attainment
	measure (instrument, scale, source)	Writing Assessment Measure (WAM), 0-28 (Dunsmuir et al., 2015).
Secondary outcome(s)	variable(s)	Writing attainment on three subtests: 1) vocabulary, 2) organisation and overall structure, and 3) ideas. Writing self-efficacy (for Year 5 only) Reading and Maths Attainment on KS2 SATs (Year 5 cohort only, at 1 year post intervention)
	measure(s) (instrument, scale, source)	Relevant subsets of the WAM, 0-4 (Dunsmuir et al., 2015). Self-Efficacy for Writing Scale, 0-100 (Bruning et al., 2013) Scaled KS2 SAT scores on reading and maths
Baseline for primary outcome	**Variable**	Writing capabilities
	measure (instrument, scale, source)	Writing Assessment Measure (WAM), 0-28 (Dunsmuir et al., 2015).
	Variable	Three subtests: 1) Vocabulary, 2) organisation and overall structure, and 3) ideas.

Baseline for secondary outcome	measure (instrument, scale, source)	Relevant subsets of the WAM, 0-4 (Dunsmuir et al., 2015).

This is a two-armed cluster randomised controlled efficacy trial, with random allocation at the school level. Cluster randomisation is necessary because Writing Roots is implemented as a whole-school programme across Years 1–6, making individual- or class-level randomisation impractical and at high risk of contamination. Whilst it is a whole-school intervention, the trial will focus on pupils in Year 2 (KS1) and Year 5 (KS2) during the 2025/26 academic year, allowing evaluation of writing skills during key stages of literacy development while minimising disruption to SATs preparation.

To be eligible for participation in this trial, schools must:

- Be state-funded mainstream primaries with full Year 1–6 provision (whole primary phase)
- Not be using Writing Roots or other Literacy Tree programmes currently or within the previous two years
- Not be participating in other whole-school EEF literacy trials
- Have no more than two mixed-year classes
- Agree to deliver the programme with fidelity if allocated to intervention

Recruitment was led by Literacy Tree through expressions of interest, with screening confirmed jointly by the delivery and evaluation teams. Schools signed memorandums of understanding (MoUs) and data sharing agreements (DSAs) prior to randomisation. A total of 110 state-funded primary schools in England were recruited, with 7,082 pupils contributing to baseline and endline data collection. We aimed for 50% of schools to be in Education Investment Areas (EIAs) to improve the relevance of findings for disadvantaged communities. At randomisation, 36.4% of recruited schools were in EIAs.

Randomisation

In this cluster randomised controlled trial (RCT), 110 schools were allocated 50:50 to the intervention and control conditions, meaning that 55 were allocated to the intervention group and 55 to control. While schools are the unit of randomisation, children are the unit of analysis, reflecting the clustered-RCT design of this evaluation. This intervention is designed to be delivered at the whole school level, so could not be randomised on the individual level while maintaining a delivery model with fidelity to Writing Root's standard delivery. Randomisation was conducted after baseline administration of the WAM to avoid bias.

Randomisation was stratified by region defined by EIA status, ensuring an equal number of schools from EIAs were included in the treatment and control conditions. Although prioritised

support for schools within EIAs was withdrawn just prior to the recruitment period[1], it is likely that at the time of recruitment and delivery, schools in EIAs were still benefiting from existing support packages and EIA status is still likely to be a relevant indicator of higher socio-economic disadvantage in the academic year 2025/26. Given the focus on understanding differential impacts of the intervention on pupils from a socio-economically disadvantaged background, stratifying on EIA status remained relevant for the purposes of the trial.

Schools were considered eligible for randomisation if: i) they had signed the MoU and DSA, ii) shared pupil data and baseline pupil scripts[2] for the WAM, and iii) that two-thirds of the scripts were deemed legible. Given the planned use of AI for marking, it was critical that scripts met a certain quality threshold on legibility to ensure they could be read by the technology. Scripts that weren't legible by AI were rescanned again by the evaluation team to improve the clarity of the handwritten text. This process continued until at least two-thirds of the scripts each school provided were considered readable.

In the protocol (Brian et al., 2025), it was stated that randomisation would take place in two blocks. There were challenges to recruitment which meant that, by the initial date set for randomisation (in the first week of July), we were below our targeted number of schools. To ensure our trail was still sufficiently powered, whilst mitigating attrition from schools who met the randomisation criteria having to endure a prolonged wait to have their treatment allocations revealed and facilitating the arrangement of INSET training for delivery schools, we set a second recruitment target at a later date in July. However, after the second randomisation date, we were still facing an underpowered trial. On balance, it was assessed that the risk to the trial from introducing an additional September recruitment drive was lower than the risk of proceeding with an underpowered trial. As a result, a third randomisation block was introduced in October 2025. Additional sensitivity analyses, outlined below, have been introduced due to the risk to validity from: a) the difference in timing of the baseline assessments (Autumn baselining compared to Summer baselining) and b) the reduction in intervention dosage received by pupils in the third randomisation block, due to the delay in training and programme delivery for Block Three schools until after the Autumn half-term holiday.

We outline in Table 2 the timing and sample sizes of each randomisation block. Note, that in the first block, two pairs of schools were recruited that requested being allocated to the same treatment status. For the purposes of the evaluation, these schools are regarded as being two unique treatment units (rather than four separate schools) and are counted below as such.

Table 2: Randomisation results

	Treatment (of which in EIA)	Control (of which in EIA)	Total (of which in EIA)

1 https://www.gov.uk/government/publications/education-investment-areas/education-investment-areas
2 The scripts were manually written by students.

Block 1 (3 July 2025)	32 (13)	32 (14)	64 (17)
Block 2 (17 July 2025)	18 (5)	18 (5)	36 (10)
Block 3 (9 October 2025)	5 (2)	5 (1)	10 (3)
Total	55 (20)	55 (20)	110 (40)

In each block, randomisation was conducted by a member of the evaluation team who was blind to the setting identities. A tailored package in Stata (*randtreat*) was used to implement the randomisation with EIA stratification. A second senior researcher at RAND Europe then checked the randomisation code and the outcomes to verify independence. The code used to randomise settings will be included in the final Evaluation Report at the conclusion of the study. A master copy of the final allocation was retained in a locked folder on RAND Europe's servers to prevent editing, and the final allocation communicated to the delivery team checked against it to ensure no edits occurred in the processing or transfer of data.

Outcomes

The primary outcome for this trial is writing attainment, measured using the Writing Assessment Measure (WAM), collected at two different time points:

i) **Baseline WAM assessment:** Baseline WAM was administered to all participating pupils, either in Summer 2025, when participating pupils were in Year 1 and Year 4, or, for those in schools in Randomisation Block 3, in Autumn 2025, when they had started Year 2 and Year 5.

ii) **Endline WAM assessment:** Endline WAM will be administer to all participating pupils in the summer term, after schools return from half-term (June – July 2026).

Administering WAM at both endline and baseline provides a consistent, reliable measure across both trial year groups, supports statistical power through strong pre/post correlations, and avoids the known limitations and gaps in available administrative data for the sampled cohorts.

WAM is a 15-minute narrative writing task in response to a standardised prompt. All pupils will be issued with the same prompt at baseline and endline, taken from Dunsmuir et al. (2014) and used in other EEF writing trials with this age group (Stone et al., 2022):

> *Imagine that you could go anywhere you wanted to on a school trip with your class and your teacher. You could go anywhere at all. Write about where you would go and what you would do.*

Pupils' scripts are scored across seven components of writing: handwriting, spelling, punctuation, grammar/sentence structure, vocabulary, organisation, and ideas. Each

component is assessed on a four-point scale, which are then summed to generate a total attainment score.

During set-up, the delivery team questioned whether the prompts validated in Dunsmuir et al. (2014) would: i) be engaging enough for children in the cohort to produce writing representative of their skills; and, ii) be equitable across schools of differing levels of social advantage. In collaboration with members of the evaluation team at the University of Leeds, they developed a new prompt which was piloted by the evaluation team prior to baseline data collection. However, the pilot results suggested that the marks for the new prompt did not approximate a normal distribution, with a clear skew, prompting concerns of a higher risk of ceiling effects under the new prompt. As a result, we have proceeded with one of the initial prompts (specified above) validated in Dunsmuir et al. (2014). The pilot is discussed in more detail in appendix A.

As part of this trial, the delivery team is developing an AI marking tool for assessing WAM scripts. This tool could be used to provide detailed and accurate assessments for each baseline script without the need for high labour resourcing, reducing the overall cost of testing pupils. This tool is being developed using a sample of 200 scripts from the baseline data of 10 schools in the main trial sample. This tool was developed in consultation with prompt-engineering specialists in RAND Europe's in-house Data Science Lab. The pilot suggests that AI marking produces similar marking distributions to human markers, with substantial rates of agreement with human markers and high levels of consistency. On the basis of these results, we plan to proceed with AI marking. Further details of how this tool was developed and validated are provided in Appendix B.

To ensure consistency in performance as we scale the AI tool, we will ensure the AI marking of endline tests is verified against human marked scripts. We will randomly select 300 endline scripts to be human marked. AI-marked scripts will be compared with the human-marked scripts, following the analysis outlined in Appendix B, to evaluate whether the AI marking tool continues to demonstrate the same level of agreement and consistency with human marking.

Secondary outcomes focus on: i) specific components of writing directly targeted by Writing Roots, and ii) writing motivation. We will examine three WAM subcomponents (vocabulary, organisation, and ideas), which reflect skills explicitly targeted by Writing Roots, and therefore may be more sensitive to programme impact in the short term. These are collected and marked according to the process outlined for the primary outcome, and are available for both year groups at both baseline and endline.

Writing self-efficacy will be measured for the Year 5 cohort only using the Self-Efficacy for Writing Scale, a 16-item Likert scale questionnaire assessing pupils' beliefs about their ability to write effectively and persist with writing tasks. The Self-Efficacy for Writing Scale will not be administered in Year 2 due to the lack of an age-appropriate, validated measure for this age group; instead, the Implementation and Process Evaluation will explore younger pupils' perceptions of writing.

A further longitudinal analysis of the long-term effect of the intervention on wider academic attainment is planned. For the Year 5 cohort, we will link pupil trial data to National Pupil Database records to analyse KS2 scaled SAT scores in reading and mathematics in Summer 2027

(one year after the intervention endline). This will allow us to identify any broader academic benefits or trade-offs associated with increased focus on writing instruction.

Sample size calculations overview

Table 3: Year 2 MDES/sample size calculations

Unadjusted									
		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)		0.141	0.163	0.164	0.190	0.153	0.167	0.182	0.206
Pre-test/ post-test correlations	level 1 (pupil)	0.70		0.70		0.70		0.70	
	level 2 (school)	0.67		0.67		0.67		0.67	
Intracluster correlations (ICCs)	level 2 (school)	0.12		0.10		0.12		0.10	
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		8		31		8	
Number of schools	intervention	65	57	65	57	55	48	55	48
	Control	65	57	65	57	55	48	55	48
	Total	130	114	130	114	110	96	110	96
Number of pupils	intervention	1950	1521	585	401	1673	1305	418	326
	Control	1950	1521	585	401	1752	1367	438	315
	Total	3900	3042	1170	802	3425	2672	856	631
Adjusted									
		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)		0.155	0.170	0.191	0.209	0.169	0.185	0.201	0.228

Pre-test/ post-test correlations	level 1 (pupil)	0.70		0.70		0.70		0.70	
	level 2 (school)	0.67		0.67		0.67		0.67	
Intracluster correlations (ICCs)	level 2 (school)	0.12		0.10		0.12		0.10	
Alpha		0.025		0.025		0.025		0.025	
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		8		31		8	
Number of schools	intervention	65	57	65	57	55	48	55	48
	Control	65	57	65	57	55	48	55	48
	Total	130	114	130	114	110	96	110	96
Number of pupils	intervention	1950	1521	585	401	1673	1305	418	326
	Control	1950	1521	585	401	1752	1367	438	315
	Total	3900	3042	1170	802	3425	2672	856	631

Table 4: Year 5 MDES/sample size calculations

Unadjusted									
		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)		0.141	0.167	0.164	0.19	0.157	0.172	0.182	0.206
Pre-test/ post-test correlations	level 1 (pupil)	0.70		0.70		0.70		0.70	
	level 2 (school)	0.67		0.67		0.67		0.67	
Intracluster correlations (ICCs)	level 2 (school)	0.13		0.10		0.13		0.10	
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	

Unadjusted									
		Protocol				**Randomisation**			
		OVERALL		**FSM**		**OVERALL**		**FSM**	
Average cluster size		30		8		33		8	
Number of schools	intervention	65	57	65	57	55	48	55	48
	Control	65	57	65	57	55	48	55	48
	Total	130	114	130	114	110	96	110	96
Number of pupils	intervention	1950	1521	585	401	1745	1361	436	340
	Control	1950	1521	585	401	1912	1491	478	372
	Total	3900	3042	1170	802	3657	2852	914	712
Adjusted									
		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)		0.161	0.175	0.191	0.209	0.174	0.190	0.201	0.228
Pre-test/ post-test correlations	level 1 (pupil)	0.70		0.70		0.70		0.70	
	level 2 (school)	0.67		0.67		0.67		0.67	
Intracluster correlations (ICCs)	level 2 (school)	0.13		0.10		0.13		0.10	
Alpha		0.025		0.025		0.025		0.025	
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		8		33		8	
Number of schools	intervention	65	57	65	57	55	48	55	48
	Control	65	57	65	57	55	48	55	48
	Total	130	114	130	114	110	96	110	96
Number of pupils	intervention	1950	1521	585	401	1745	1361	436	340
	Control	1950	1521	585	401	1912	1491	478	372
	Total	3900	3042	1170	802	3657	2852	914	712

The trial has been designed to provide adequate statistical power to detect a meaningful improvement in writing attainment outcomes for pupils in Year 2 and Year 5. Sample size calculations were conducted using PowerUp! software and informed by assumptions derived from previous EEF literacy trials and the existing evidence base on the WAM.

We assume pre-test/post-test correlations of 0.70 at pupil level and 0.67 at school level, which are conservative compared with test–retest evidence for the WAM (Dunsmuir et al., 2015) and broadly in line with similar measures in other EEF trials, such as Helping Handwriting Shine (Stone et al., 2022). Intracluster correlation coefficients (ICCs) were set at 0.12 for Year 2 and 0.13 for Year 5 for all pupils, and 0.10 for FSM pupils, taking into consideration the recommended ICC of 0.10 provided in Singh et al, (2023) and our own review of EEF writing trials which produces average ICCs for Year 5 of 0.17 and for Year 2 of 0.12. At protocol stage, school clusters were assumed to contribute 30 pupils per cohort on average, including eight pupils eligible for FSM.

Our initial recruitment target was a total of 130 schools with 65 to be assigned to the intervention group and 65 to control, with an assumed population of approximately 3,900 pupils per year group at baseline. This sampling is powered to assess impacts separately in both Year 2 and Year 5. However, we failed to meet this recruitment target, motivating the block randomisation structure described in the Randomisation section. We succeeded in recruiting 110 schools before the randomisation stage, and maintained 1:1 randomisation (55 in intervention; 55 in control). The average cluster size was slightly higher than assumed at protocol stage for the overall sample, with an average of 31 in Year 2 and 33 in Year 5.

Tables 3 and 4 display the Minimum Detectable Effect Sizes (MDES) calculations and associated assumptions. As there are two primary research questions we present results for the MDES both unadjusted and adjusted for multiple hypotheses. The unadjusted MDES under main assumptions, at randomisation, are 0.153 for Year 2 and 0.157 for Year 5 (see Tables 3 and 4 respectively). For the FSM subgroup, the MDES at randomisation are 0.182 for both Year 2 and Year 5. These values are higher than the MDES calculated at the protocol stage, driven by lower recruitment levels. Nevertheless, under the assumptions outlined here, they fall below the conventional 0.20 benchmark associated with strong confidence in causal evidence, and below the effect sizes typically seen in writing trials. When adjusting for multiple hypotheses, the MDES rises to 0.169 for year 2 and 0.174 for year 5. The adjusted MDES for FSM analysis is 0.201, which is slightly above the 0.20 benchmark.

We maintain attrition assumptions reflecting typical levels in primary literacy trials: 12% school attrition and 22% pupil attrition (see Table 5). Table 6 below shows the MDES after these assumptions for the overall sample. Under combined attrition, unadjusted MDES values increase modestly to 0.167 for year 2 and 0.172 for year 5, and adjusted MDES values increase to 0.185 for year 2 and 0.190 for year 5. The MDES for the FSM analysis rises to 0.206 or 0.228 when adjusted, taking it above the conventional threshold for a well-powered study. Although the trial is not explicitly powered with FSM pupils as the primary population, the study maintains acceptable MDES levels for subgroup analysis, provided attrition is limited. Under these conditions, robust exploration of differential effects for disadvantaged pupils will be possible.

Table 5: Assumptions over attrition

	School level	Pupil level
Expected attrition (%)	12%	22%

Table 6: MDES after attrition

Unadjusted			
	Pupil attrition only	**School attrition only**	**Pupil and school attrition**
Year 2	0.157	0.163	0.167
Year 5	0.161	0.168	0.172
Adjusted			
	Pupil attrition only	**School attrition only**	**Pupil and school attrition**
Year 2	0.173	0.180	0.185
Year 5	0.178	0.186	0.190

Analysis

The statistical analysis proposed follows the most recent revised EEF Statistical Analysis Guidance (2022) available here.

Primary outcome analysis

As specified in the protocol (Brian et al., 2025) this efficacy trial has two key research questions

1. What is the impact of Writing Roots on writing outcomes for pupils in Year 2 as measured by the Writing Assessment Measure (WAM)?
2. What is the impact of Writing Roots on writing outcomes for pupils in Year 5 as measured by the WAM?

To address these questions, we will estimate a multilevel model of writing attainment (equation 1) using WAM (see Design Overview section), the primary outcome for this efficacy trial. This analysis will be conducted on an Intention To Treat (ITT) basis. Under this approach, schools will be analysed as randomised, with all randomised schools and tested pupils being included in the analysis according to their assigned treatment regardless of the treatment actually received, or any deviations in the delivery of the intervention. The ITT approach is inherently conservative as it captures the averaged effect of offering the intervention, regardless of whether the participants complied with assignment. This principle is key in ensuring an unbiased analysis of intervention effects and is in line with the EEF's guidance (EEF, 2022).

More specifically, using multi-level modelling (MLM), we will estimate a two-level random intercept model for each year group. The two-level model, with the first level the unit of analysis (the pupil) and the second level the unit of randomisation (the school), reflects the trial design and nested nature of the data, as recommended by the statistical guidance (EEF, 2022). It appropriately clusters the error term at the unit of randomisation to ensure appropriate, unbiased confidence intervals are estimated. The two-level random-intercept model estimates a single average treatment effect on writing attainment, whilst allowing for setting-specific variation in mean attainment.

The impact will be estimated using the model outlined below in equation 1. Equation 1 is known as a 'random intercepts' model because $\beta_{\{0j\}} = \beta_0 + u_j$, the setting-specific intercept for school j, is random, with an assumed distribution ($\beta_{0j} \sim_{i.i.d} N(\beta_0, \sigma_u^2)$). The model will control for baseline attainment, as measured by pre-test WAM scores, and estimate fixed effects for the stratification variable (EIA) at the setting level.

$$Y_{ij} = \beta_0 + \tau WR_j + \beta_1 Z_j + \beta_2 X_{ij} + u_j + e_{ij} \quad (1)$$

Where:

Y_{ij} represents the post-test (endline) WAM total score for pupil i in school j.

β_0 is the cluster-level coefficient for the slope of a predictor on writing skills

WR_j, a binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group

Z_j represents the stratifying variable of EIA-status used in the randomisation for school j

X_{ij} represents the baseline WAM score for student i in school j

u_j represents setting-level residuals

e_{ij} represents child-level residuals.

The coefficient τ is the main outcome of interest, as an estimate of the conditional effect of treatment on endline writing attainment, as measured by the WAM. We use τ to calculate a standardised Hedges' g effect size (see Calculation of Effect Sizes section below).

This analysis will be conducted separately for Year 2 and Year 5 pupils. KS1 and KS2 pupils are at different development stages, impacting their learning and response to interventions, necessitating separate analyses (Goswami, 2015; EEF, 2020). Separate analyses will also improve statistical precision and power by reducing variability. Distinct analyses allow for clearer interpretation of results, making it easier to identify specific effects for each year group.

Given we have two cohorts in the primary analysis, and so are testing two hypotheses on the primary outcome, we will employ a Romano-Wolf correction to estimated effect sizes, to statistically correct for over-rejection of null hypotheses under multiple hypothesis testing. We expect there to be some degree of correlation between the two samples, considering they are clustered in the same school. In such circumstances, we prefer to apply the Romano-Wolf correction, to take into account the dependent nature of the test statistics and provide a strong control against the family-wise error rate (Clarke et al., 2019), as recommended in the EEF's statistical analysis guidance (Education Endowment Foundation, 2022).

All analyses will be done in Stata, version *18* and above, or R using the *eefanalytics* package (Vallis et al., 2021).

Secondary outcome analysis

The efficacy trial will seek to answer the following secondary research questions (Brian et al., 2025):

3. What is the impact of Writing Roots on writing outcomes for pupils in Year 2, as measured by the following WAM subtests:
    a. Vocabulary
    b. Organisation and overall structure
    c. Ideas
4. What is the impact of Writing Roots on writing outcomes for pupils in Year 5, as measured by the following WAM subtests:
    a. Vocabulary
    b. Organisation and overall structure
    c. Ideas
5. What is the impact of Writing Roots on writing self-efficacy for pupils in Year 5, as measured by the Self-Efficacy for Writing Scale?

As discussed in the Outcome Measures section, this trial will use a number of secondary outcome measures. Three of these are subscales of the WAM test, which we will use to assess progress in vocabulary, structure, and ideas - the elements of writing that are identified as fundamental short-term outcomes in the Writing Roots theory of change available in the protocol (Brian et al., 2025). The secondary analysis using these outcome measures will answer RQ3 and RQ4.

The secondary analysis using WAM subsets will be assessed using the same multi-level modelling approach outlined in the 'Primary outcome analysis' section. As in the primary outcome analysis, the model accounts for baseline achievement in each of the WAM subscales, using pre-test scores for the relevant subscale, and will estimate fixed effects for the EIA stratification variable at the school level. Analysis for each will follow the specification of equation 2. Each subscale is modelled independently of the others, so the below model is repeated three times, once for each of the subscales

$$Y_{ij} = \beta_0 + \tau WR_j + \beta_1 Z_j + \beta_2 X_{ij} + u_j + e_{ij} (2)$$

Where:

Y_{ij} represents the post-test (endline) subscale score (Vocabulary, Organisation and Overall Structure, or Ideas) for pupil i in school j.

β_0 is the cluster-level coefficient for the slope of a predictor on writing skills

WR_j, a binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group

Z_j represents the stratifying variable of EIA-status used in the randomisation for school j

X_{ij} represents the baseline subscale score (Vocabulary, Organisation and Overall Structure, or Ideas) for student i in school j

u_j represents setting-level residuals

e_{ij} represents child-level residuals.

The coefficient τ is the main outcome of interest. We will use τ to calculate a standardised Hedges' g effect size (see Calculation of Effect Sizes) of the intervention of each of the subscales.

For RQ5 We will be conducting two secondary analyses using the Self-Efficacy for Writing Scale as the outcome of interest. The first will follow the specification outlined in equation (2) using the Self-Efficacy for Writing Scale as the endline outcome (Y_{ij}) and using the WAM total score as the baseline measure, accounting for pupil-level differences in ability (X_{ij}).

This will be accompanied by a sensitivity analysis where the effect of Writing Roots on self-efficacy will be evaluated using an endline only model, using the following specification:

$$Y_{ij} = \beta_0 + WR_j\tau + Z_j\beta_1 + u_j + e_{ij} (3)$$

Where:

Y_{ij} is the Self-Efficacy for Writing Scale outcome at endline

β_0 is the cluster-level coefficient for the slope of a predictor on writing self-efficacy

WR_j, a binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group

Z_j represents the stratifying variable of EIA-status used in the randomisation for school *j*

u_j represents setting-level residuals

e_{ij} represents child-level residuals.

Note that in this model, pupil-level characteristics are not controlled for as they are in equation (1) or (2). The coefficient τ on the Writing Roots dummy (WR_j), still signifies the main outcome of interest: the estimated impact of the Writing Roots programme on writing self-efficacy.

Given we have multiple secondary outcomes, we will employ a Romano-Wolf correction to estimated effect sizes, to statistically correct for over-rejection of null hypotheses under multiple hypothesis testing. Dunsmuir et al. (2014) shows that there is substantial correlation between different subscales of the WAM (between the three subscales used in the secondary analysis, the inter-item Pearson's *r* coefficient ranges from 0.46-0.8 and all are significant at the 0.01 level) and with the total score. Measures of writing self-efficacy also display significant relationships with measures of writing ability (Bruning et al., 2013). In such circumstances, the Bonferroni correction might be too conservative and lead to an over-correction. For this reason, we will apply the Romano-Wolf correction in secondary outcomes analyses. To ensure alignment with the primary analysis, we will adjust for all eight hypotheses (each of the four hypotheses for each year group). This correction will take into account the dependent nature of the test statistics and will provide a strong control against the family-wise error rate (Clarke et al., 2019), as recommended in the EEF's statistical analysis guidance (Education Endowment Foundation, 2022).

All analyses will be done in R or Stata, 18 or above, making use of the *eefanalytics* package where possible. Multiple hypothesis testing corrections will be made using the *rwolf2* package in Stata (Clarke, 2021) or *crctStepdown* in R (Watson, 2024) to correct for this.

Subgroup analyses

As specified in the protocol the trial will explore the following research question:

6. Are the impacts differential for pupils eligible for free school meals (FSM), pupils with SEND, or those with English as an additional language (EAL)?

Sub-group analysis will be performed for the primary outcome in Year 2 and Year 5 according to EEF's evaluation guidance (EEF, 2022). Due to greater reliability over primary-collected indicators, we will be using the NPD to categorise pupils into their subgroups. The subgroups are defined as follows:

- **FSM eligibility**: we will identify pupils' eligibility for FSM using the binary *EVERFSM_6_P* variable from the NPD. Pupils will be classed as eligible for FSM if *EVERFSM_6_P* takes on a value of one.

- **SEND:** we will identify SEND pupils using the variable *SENprovisionMajor*. This is a categorical variable, with pupils classified into one of four different categories: no SEND, SEND without an Education, Health, and Care Plan (EHCP), SEND with an EHCP and pupils who are not given a SEND classification. For the purposes of the subgroup analysis this be reduced to a binary variable, as children without a SEND classification will be removed from the subgroup analysis and children with an EHCP are not included in in data collection or analysis (they will still receive the intervention). The exclusion of children with an EHCP from analysis means that the results of the subgroup analysis will not be generalisable to all SEND pupils.

- **EAL:** we will identify EAL pupils using the variable *LanguageGroupMajor*. This is a categorical variable, with pupils classified into one of three different categories: English, a language other than English, or no primary language classification. EAL status will be reduced to a binary variable, with pupils without a primary language classification removed from the subgroup analysis.

There is a chance that the subgroup analysis will be under-powered. Given the trial has been powered to detect a moderate effect on the overall sample, there is a risk that all sub-group analysis may be under-powered due to the smaller sample sizes. At randomisation, we are sufficiently powered to detect a moderate effect on the FSM subgroup; however, depending on attrition rates, we may still be under-powered at analysis.

Analysis of the subgroups will use two approaches, as suggested in EEF analysis guidance (EEF, 2022). The first will run the primary model given in equation 1 on the FSM, SEND, and EAL subgroups only, following the procedure outlined in the above section on 'Primary outcome analysis'. The effect size of Writing Roots on each subgroup will be calculated and reported as outlined in the 'Calculation of Effect Sizes' section.

For each subgroup, we then estimate the treatment effect of Writing Roots using a second approach, which makes use of the entire sample. This second model estimates subgroup treatment effects using an interaction model, here using FSM as an example:

$$Y_{ij} = \beta_0 + \tau WR_j + \beta_1 FSM_i + \beta_2 (FSM_i * WR_j) + \beta_3 Z_j + \beta_4 X_{ij} + u_j + e_{ij} \quad (4)$$

This is the same model specification as in equation 1, with the addition of the FSM_i indicator of disadvantage and an interaction term combining FSM eligibility and treatment allocation ($FSM_i * WR_j$). The primary coefficient of interest in the interaction model is β_2, which can be interpreted as the additional treatment effect experienced by pupils from disadvantage background: a positive β_2 is indicative of a treatment acting as a 'gap-closer' and a negative β_2 indicative of treatment acting as a 'gap-widener'. For the SEND and EAL subgroups, equation 4 is repeated using SEND and EAL binary indicators instead of the FSM_i indicator.

The treatment effect size from the interaction models is calculated according to the following formula:

$$\frac{\tau WR_j + \beta_2(WR_j * FSM_i)}{sd}$$

The coefficients in the numerator come directly from equation 4, and the standard deviation used in the denominator is the unconditional standard deviation of the FSM sub-sample (both treatment and control).

Sensitivity analysis

Adjusting for differing baseline test timings

As outlined in the Design Overview, recruitment issues meant that we deviated from the initial randomisation plan laid out in the protocol, instead randomising schools in three different 'batches'. The last of these batches had their randomised allocation revealed to them in October, after intervention schools who had their allocations revealed to them in the summer term had already started to receive the treatment. This presents two issues for the trial: i) baselining for the third batch took place in the Autumn term instead of the summer term, which could reduce baseline scores for these schools as a result of summer learning loss (EEF, 2020) and ii) intervention schools from this batch received a reduced dose of the Writing Roots programme, as they missed the first half-term of delivery. This means that the primary analysis could be biased by the inclusion of these schools. To ensure these are the only risks the Autumn recruitment poses to the trials, we will as a robustness check cross-tabulate all baseline and demographic variables for the Summer and Autumn schools and check for imbalance.

Our sensitivity analysis proposes treating the summer-baselined schools (those in randomisation batch 1 and 2) as a subgroup, and thus follows the subgroup analysis outlined in the previous section. We will repeat the primary analysis model, outlined in equation 1, for the summer-baselined subsample only (i.e., excluding the ten schools in randomisation batch 3). Following EEF guidance, we will also estimate an interaction model, following the specification in equation 4, on the full sample and estimate the effect size according to the equation outlined in the subgroup analysis section. We will compare these estimated effect sizes to those estimated in the primary analysis to understand whether the inclusion of autumn-baselined schools results in a bias of the effect size. If this sensitivity analysis suggests substantial bias has been introduced into the primary outcome model from the inclusion of these schools, we will caveat the primary findings appropriately.

As outlined above, schools randomised in October also receive a lower dosage. As such, we will make adjustments to our dosage and compliance analysis to reflect this. This is outlined in the Compliance Analysis Section below, but effectively caps the number of lesson plans they can use to 10 out of 12, to reflect the loss of half a term of intervention delivery.

Adjusting for differences in recruitment for small schools

Schools with 15 or fewer pupils per year group were initially only considered for recruitment to the trial as part of paired units with schools with similar management, However, as recruitment

progressed it became apparent that some small schools had been individually recruited. As there were no explicit eligibility criteria in the MoU restricting recruitment by size, these schools were allowed to proceed with randomisation.

The primary analysis will analyse schools as randomised. Where schools were recruited and randomised as a paired school, the schools will be analysed as a single unit, rather than two separate schools, given they were randomised as a single unit. Where schools were recruited and randomised individually, the schools will analytically be kept as individual units, regardless of their size.

However, given the inconsistencies in the recruitment approach to small schools, we will conduct two sensitivity analyses:

i) Re-run the model with all paired units analysed as if there were not paired. This means each school will be treated as a single unit, regardless of whether or not they were recruited and randomised as a pair. This treats all schools equally, regardless of size.

ii) Re-run the model dropping all schools with fewer than 15 students in each form, whether paired or unpaired. This would more substantially reduce power, but ensures that we remove small schools who were treated differently during recruitment.

Exploratory analysis

Analysis of effect on pupils with varying pre-intervention attainment levels

Given previous evaluations have noted marked differences between low and high attainers (Slavin et al., 2019), we will explore whether there are differential impacts for learners from different attainment quartiles at baseline. We will generate two different binary indicators:

- **low attainers:** takes a value of 1 if the baseline WAM scores are in the lowest quartile for their cohort and 0 otherwise
- **high attainers:** takes a value of 1 if the baseline WAM scores are in the highest quartile for their cohort and 0 otherwise.

We will model effectiveness of Writing Roots on different attainment levels as per the models outlined in the subgroup analysis section. Following EEF analytical guidance, we will estimate effect sizes both by estimating the primary model (equation 1) on the attainment subgroups only, and by estimating an interaction model (equation 4) using the full sample. All effect sizes will be calculated as outlined in the subgroup analysis section.

Analysis of a combined-cohort model

Given this is a whole-school intervention, we will seek to estimate the impact of Writing Roots at the school level by combining data from Year 2 and Year 5 cohorts. To estimate the overall school-level effect of the Writing Roots intervention across both Year 2 and Year 5 cohorts, we will implement a three-level hierarchical linear model. This approach follows EEF guidance (EEF,

2022) in reflecting the nested structure of the trial, in which pupils (level 1) are nested within year groups (level 2), which are in turn nested within schools (level 3). This approach is modelled as follows:

$$Y_{ikj} = \beta_0 + WR_j\beta_1 + C_{kj}\beta_2 + (WR_j * C_{kj})\beta_3 + Z_j\beta_4 + X_{ij}\beta_5 + u_j + v_{jk} + e_{ikj} \quad (5)$$

Where:

Y_{ikj} represents the post-test (endline) WAM total score for pupil i in cohort *k* school j.

β_0 is the cluster-level coefficient on writing skills

WR_j, a binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group

C_{kj} , a binary indicator indicating cohort, taking on value of 1 if the pupil is in Year 5 and 0 if the pupil is in Year 2

$(WR_j * C_{kj})$, an interaction term that allows for the effect of the intervention to vary by cohort

Z_j represents setting-level characteristics for setting j—specifically the stratifying variable of EIA-status used in the randomisation

X_{ikj} represents child-level characteristics for child i in setting j—specifically the baseline WAM score

u_j represents setting-level residuals

v_j represents cohort-level residuals

e_{ij} represents pupil-level residuals.

This model in equation 5 allows for the estimation of both a pooled effect and a cohort-specific effect, while accounting for the hierarchical structure of the data and potential correlation between cohort-specific outcomes within schools, and using the full statistical power of the trial. In accordance with EEF guidance on the use of interaction models, we will calculate the cohort-specific treatment effect sizes from the interaction model and compare these with the effect sizes for each cohort generated in the primary analysis.

In the protocol, we suggested this would either be explored through pooling the data from the cohorts or combining the separate primary analyses into a single aggregated effect size using meta-analytic techniques. The three-level model is preferred as it aligns with the nested structure of the trial and provides a statistically robust framework for estimating the school-level impact of the intervention. In line with alternative approaches, such as fixed-effect meta-analysis, this model can still estimate cohort-specific fixed treatment effects through the interaction term. However, unlike alternative approaches, it models the full nested structure of year-group cohorts within schools, rather than assuming cohorts are independent samples. We

will report the regression output, variance-covariance matrix, and intracluster correlations to allow us to analyse both the fixed effects and independence of cohorts assumptions.

If the model encounters convergence issues or if the cohort-within-school variance component is negligible, we may consider simplifying to a two-level hierarchical model (pupils nested in schools) which broadly follows the specification in Equation 5 without the inclusion of cohort random effects, v_{kj}. This approach retains the ability to test for differential effects across cohorts while reducing the complexity of the random effects structure.

The *eefanalytics* package does not accommodate three-level hierarchical models, therefore for this analysis, we will use *lmer* package in R or the *mixed* command in Stata.

Longitudinal analysis[3]

To understand the longer-term effects of Writing Roots on attainment, we will conduct follow-up analysis of attainment in reading and mathematics one year after intervention delivery for the Year 5 pupils (when pupils are at the end of KS2). We will examine the following research question:

7. Does participation in Writing Roots lead to long-term improvements in reading and maths attainment at KS2 for the Year 5 cohort?

There is potential for spillover effects on participants on other learning outcomes (both positive and negative). For example, deep engagement with texts as part of the programme could improve reading skills. Improvements in reading and writing skills could improve broader attainment in all subjects; however, there could also be negative unintended consequences, whereby a whole-school curriculum focus on one component (writing) could decrease teaching time and focus on other areas (such as mathematics). We will use data in the NPD to capture the following KS2 outcomes:

- **KS2 Reading attainment:** we will use scaled score in reading in the SATs at the end of KS2, *KS2_READSCORE*
- **KS2 Mathematics attainment:** we will use scaled score in mathematics in the SATs at the end of KS2, *KS2_MATSCORE*. Pupils will be analysed as randomised, in their original clustering and sub-groups, even if they have moved schools or may be classified differently in the NPD at the point of longitudinal analysis. For example, if pupils did not have an EHCP in place by endline analysis, they will be retained in longitudinal follow-up even if they receive an EHCP in the intervening year.

This analysis will be pre-specified in an update to this SAP after the resolution of some outstanding queries relating to trial school's access to Writing Roots in the next academic year (Year 6 SATs year for the current Year 5 cohort).

[3] Please see the longitudinal analysis guidance.

Imbalance at baseline

A well-conducted randomisation should create groups that are equivalent on observables at baseline, with any imbalance at baseline occurring by chance (Glennerster & Takavarasha, 2013). To check for imbalance at baseline after randomisation, baseline equivalence testing will be conducted at the school and pupil level. At the school level, we will examine the balance over EIA distribution, Ofsted ratings and proportion eligible for FSM. These will evaluate the distribution of each characteristic between the control and intervention groups. At the pupil level, balance will be assessed over: gender, FSM status, SEND status, EAL status and attainment in all baseline tests. As recommended in the EEF statistical analysis guidance (EEF, 2022), we will report differences in pupil-level pre-tests as effect sizes with accompanying confidence intervals. The balance at baseline will be on the sample as randomised. Should there be imbalance on baseline characteristics, we will repeat all primary and secondary analysis controlling for any covariates that demonstrate imbalance at baseline as a robustness check.

Missing data

Missing data can arise from pupil non-response to outcome testing (i.e., refusing to participate), measurement attrition of participants at school and pupil levels (either due to withdrawal or, at the pupil-level, absences), or test administration errors. Unfortunately, non-random missingness can introduce bias into the ITT approach outlined in the primary analysis section above.

For all primary, secondary and subgroup analysis, we will report the extent of missing data through cross-tabulations. In line with EEF guidance (EEF, 2022), if there is less than 5% missingness overall, we propose to only carry out a complete-case analysis, under the assumption that the data are missing completely at random (MCAR). If the missing data exceed 5% of the sample as randomised, our approach will depend on the pattern of missingness observed. For the primary outcome measure, we will additionally analyse the pattern of missingness and, if suitable, perform a multiple imputation (MI) analysis, as recommended in the EEF guidance (EEF, 2022).

To assess whether there are systematic differences or a clear pattern in missingness, we will model missingness at follow-up (defined as pupils with missing primary outcome data at endline) as a function of covariates available at baseline. The analysis model for this approach will mirror the multilevel level model specified in equation 1, with pupils clustered at the setting level. In this instance, however, given the outcome will be a binary variable identifying missingness (where 1=missing; 0=complete), we will use a multilevel mixed-effects logistic regression model. All available setting-level and individual-level variables will be used in this model, not just those pre-specified in equation 1. On the basis of the logistic 'drop-out' model, we will assess the pattern of missingness and take the following approach:

- If the missing data pattern appears to be unrelated to any observables, and or possible unobservable variables (for instance, solely due to pupil absences), we will presume that the data are MCAR and proceed with an analysis based only on complete cases.

- If there is evidence that missingness is correlated with observable covariates, then data is likely at least missing at random (MAR) and a complete-case analysis will be biased. Depending on the pattern of missingness, MI may be warranted. Both full-information maximum likelihood (FIML) and MI have been shown to be broadly equivalent (Lee & Shi, 2021). We will follow the guidelines for MI recommended in Jakobsen et al. (2017).

- If missing data seems likely to depend on unobserved variables, even after controlling for all observable covariates, then the data is likely missing not at random (MNAR). Complete case analysis is likely to be biased, but MI will not be sufficient to correct for this. Instead, sensitivity analysis will be carried out and reported alongside the headline impact estimates.

Compliance analysis

As the ITT approach is inherently conservative, capturing the average treatment effect of offering the intervention, we also propose to look at treatment effects in the presence of compliance. In a situation of imperfect compliance, whereby not all participating setting are deemed perfectly compliant, EEF analytical guidance recommends undertaking a complier average causal effect (CACE) analysis. We will use a two-stage least squares (2SLS) estimation with random group allocation serving as the instrumental variable (IV) for the compliance indicator following the EEF guidance (EEF, 2022). The CACE analysis rests on four main assumptions (Baiocchi, Cheng and Small, 2014; Angrist, Imbens and Rubin, 1996; Labrecque and Swanson, 2018; Lousdale, 2018; Raudenbush and Bloom, 2015):

1. *Exogeneity or independence*: There are no confounding effects of treatment assignment that determine writing attainment.

2. *Relevance:* Assignment into the treatment arm, and thereby being offered Writing Roots, affects compliance with Writing Roots.

3. *Exclusion restriction*: Being offered the intervention has no direct effect on outcomes except through complying with the intervention, as measured by the compliance metric.

4. *Monotonicity*[4]: Being randomised into the treatment arm does not decrease the likelihood of receiving Writing Roots. That is, there are no defiers – schools who are more likely to pursue Writing Roots if they are randomised into the control arm than the treatment arm.

Random assignment into treatment and control arms ensures the *exogeneity* assumption is satisfied. Similarly, control schools are barred from accessing Writing Roots during the trial period, even if they leave the trial[5], ensuring that the *montonicity* assumption is satisfied (being

[4] An alternative assumption to monotonicity is that of effect homogeneity. Under homogeneity, CACE can estimate the average treatment effect (ATE), rather than just the local average treatment effect (LATE). However, homogeneity is a stronger assumption that is unlikely to hold. As such, we proceed under the assumption of monotonicity and acknowledge that this means we can only measure LATE.

[5] There is one instance in which schools were given special dispensation to leave the trial post-randomisation but still receive Writing Roots – three schools signed up and it was only revealed post-randomisation that they were not only part of the same multi-academy trust, but they shared the same curriculum lead and all resources and INSET

randomised into the treatment arm will increase the likelihood of receiving Writing Roots, as the likelihood of receiving Writing Roots for control schools is necessarily zero). The *relevance* assumption will be empirically tested.

However, after publication of the protocol, concerns arose over whether the defined compliance metric would ensure the *exclusion restriction* would be satisfied. In the protocol, it was stated that compliance will be a composite binary measure formed on various data points that will come from Literacy Tree, as outlined in Table 7.

Table 7: Protocol compliance metrics

Metric	Source of data	Threshold for compliance
Lesson plan downloads	Number of downloads by a specific user captured by the Literacy Tree website (backend data source) provided by Literacy Tree	At least 9 out of 12 Writing Roots lesson plans downloaded for Year 2 and at least 9 out of 12 Writing Roots lesson plans downloaded for Year 5.
Attendance at INSET training	Training attendance logs, or (where catch-up is needed) viewing statistics of individual training catch up videos from Literacy Tree website. Both provided by Literacy Tree.	All teachers attend the INSET day or watch at least 70% of an INSET catch-up video
Attendance at Teach Through Text sessions	Attendance logs from Literacy Tree.	All teachers attend two out of three Teach Through a Text sessions

This would have been a binary measure whereby schools would have been deemed compliant if they met all of these conditions, and non-compliant if they did not. The dichotomisation of a multi-dimensional conceptualisation of compliance poses concerns for the exclusion restriction, however, as it would mean that partially-compliant schools would be deemed non-compliant. The exclusion restriction requires that treatment allocation cannot influence writing performance outcome other than through our defined compliance metric. If partially compliant schools are still effective at improving writing attainment, then treatment allocation can still effect the final outcome, directly violating the exclusion restriction. After extensive discussion with the delivery team, and research team attendance at INSET and Teach Through a Text sessions, it became evident that partially compliant schools may still positively effect writing attainment. For instance, schools who only attended one Teach Through a Text session but

training across schools, making it impossible for the schools to be randomised to different treatment arms. These schools were asked to leave the trial, but were given the option to pay for and receive Writing Roots, given the impossibility for the schools to comply with their randomised assignments.

taught writing according to the highly detailed lesson plans may still be effective in improving writing attainment. Violations of the exclusion restriction will result in a downward bias in CACE. Whilst our concern is specific to the definition of compliance published in the protocol, we note that more generally, dichotomisation of compliance for use in 2SLS has been shown empirically to downward bias the CACE (Merrill and McClure, 2015).

This threat to the validity of the CACE analysis motivated a reappraisal of the approach outlined in the protocol, and instead prompted a move to continuous compliance metrics. We explored using a dichotomised compliance metric where all partially-compliant schools would be deemed compliant (i.e., schools would be deemed compliant if they attended any training or downloaded any lesson plans) to avoid violations of the *exclusion* restriction. Whilst this ensures that only truly non-compliant schools would be classified as such, there were concerns that this definition of compliance would not be meaningful. In line with other writing trials, such as the Rehearsal Room Writing trial (Flemons et al., 2025), we switch to continuous compliance metrics instead, which reduces the risk of violations of the exclusion restriction. This outlined approach was agreed upon with the delivery team and EEF.

We propose a single continuous compliance metric, that measures compliance on a scale between 0 to 1, from non-compliance to full optimal compliance. Optimal compliance is still defined as attendance at all training and downloading the necessary lesson plans for delivery, as initially proposed. However, compliance is now measured on a term-by-term basis. Compliance for each term is defined as attendance at the relevant Teach-Through-a-Text training and downloading of the term's lesson plans. For most schools, the compliance metric can take on four values: 0 (zero terms compliance), 0.33 (one term of compliance), 0.66 (two terms of compliance) and 1 (three terms of compliance). For the Autumn schools, where the first half term was missed, the maximum compliance score is 0.83. Compliance will be collected at the year-group level, based on the training attendance and downloads of the teachers in that year group. As a result, year groups within the same school can have varying levels of compliance.

So long as there is variance in compliance, we will undertake CACE analysis using 2SLS estimation to recover the local average treatment effect (LATE) of attending a compliant setting on writing outcomes. The first stage of this 2SLS approach will estimate the extent to which the assignment to the intervention affects the likelihood of the setting taking up the treatment (the first stage regresses treatment assignment on compliance). This will estimate a compliance rate and will be estimated at the teacher-level, with each cohort estimated separately in line with the approach to primary analysis, using the following equation:

$$Y_{kj} = \beta_0 + WR_j\beta_1 + Z_j\beta_2 + u_{kj}$$

Y_{kj} is the compliance score of the teacher of year group *k* in school *j*

WR_j is a binary indicator equalling 0 if the school is assigned to the control group and 1 if it assigned to the treatment group

Z_j represents setting-level characteristics for setting j—specifically the stratifying variable of EIA-status used in the randomisation

u_{kj} the teacher-level residuals of the first stage

Results for the first stage, and the associated F stat, will be reported. The second stage of the IV estimation predicts the outcome as a function of all covariates included in equation 1 (see Analysis - Primary Outcomes) but substitutes the treatment indicator (WR_j in equation 1) with the compliance rate estimated in the first regression (Angrist &Krueger, 1991; Angrist, 2006). Due to ease of estimation, we will use an OLS IV approach to estimate CACE, clustering the errors at the school level. Given each year-group will be estimated separately, we do not need to cluster at cohort level as well. This does not mimic exactly the multi-level hierarchical analysis employed throughout the rest of the analysis, but still controls for intracluster correlations at the setting level to ensure appropriate standard errors and confidence intervals are used. This model will be estimated for the primary outcome measure only.

However, whilst switching to a continuous variable overcomes the risk posed by dichotomisation of compliance, there remains concerns over the multi-dimensional conceptualisation of compliance. Acknowledging that compliance is multi-dimensional means there remains possibilities that the *exclusion restriction* could still be violated. For instance, schools with high lesson-plan use but low training attendance may still pose a violation to the exclusion restriction when estimating CACE, as there is a still a pathway by which treatment assignment may effect writing attainment without affecting the compliance metric (i.e., through use of lesson plans without attending training). The same holds for schools with high training attendance but low use of lesson plans. Unless there is limited variance in one of the compliance metrics (e.g., all schools download the same number of lesson plans or all schools attend the same number of Teach Through a Text sessions), violations of the *exclusion restriction* may still occur. We will report the full distribution of each compliance dimension and discuss the implications of this, with respect to the satisfaction of the *exclusion restriction* and the validity of a 2SLS approach. Any estimates will be caveated with a discussion of the risk of a downward bias in CACE.

Compliance analysis will be undertaken in R or Stata, version 18 or higher.

Intra-cluster correlations (ICCs)

The ICC is a crucial metric for trials involving clusters. It quantifies the fraction of variance in a specific outcome attributable to differences between clusters (such as settings), rather than variance occurring within these clusters. This value is derived from the ICC observed in previous EEF trials.

In the final report, we will present ICCs as they were at the protocol stage, at the time of randomisation, and at the analysis stage. The ICC at the analysis stage will focus on the primary outcome measure. Its calculation will involve two approaches:

(i) using the model corresponding to equation 1, and

(ii) employing a model akin to that in equation 1 but without any covariates.

This second model accounts for the clustering of pupils in schools and is referred to as the 'empty model'.

ICCs will be estimated using Stata's *estat icc* command, or the *performance* package in R.

Effect size calculation

The effect size (Hedge's g) will be standardised using unconditional variance in the denominator and confidence intervals will be reported to communicate statistical uncertainty in line with EEF guidance. This will tell us the average effect of the intervention on pupil's writing outcomes in treatment settings compared to those in control settings.

As outlined in the Analysis section, unless otherwise stated, we will calculate effect sizes (hereafter ES) for cluster-randomised trials using Hedges g as outlined in the EEF statistical analysis guidance (EEFoundation, 2022):

Where:

$$ES = \frac{\left(\bar{Y}_T - \bar{Y}_C\right)_{adjusted}}{sd}$$

$\left(\bar{Y}_T - \bar{Y}_C\right)_{adjusted}$ = mean difference between the intervention and control group adjusted for baseline test score and other stratification variables

sd = estimate of the pooled unconditional standard deviation.

The pooled unconditional standard deviation is the weighted average of standard deviations of treatment and control (Coe, 2002). The pooled unconditional standard deviation across the two trial arms is used in the denominator, as we assume the standard deviations of both the treatment and control groups are drawn from the same underlying population distribution. If there is cause to question this assumption at the analysis stage, we will use the unconditional standard deviation of the control group, in line with EEF guidance.

ES will be computed for each of the estimated models and reported alongside their 95% confidence interval (CI). The ES will then be converted into months of progress (for attainment measures) to facilitate interpretability.

All ES will be estimated using the *eefanalytics* package. If there is evidence of non-normality in residuals of any models, we will report non-parametric bootstrapped confidence intervals instead, as per the *eefanalytics* package.

References

Anders, J. Shure, N., Wyse, D., Bohling, K., Sutherland, A., Barnard, M. & Frerichs, J. (2021). Learning About Culture. Overarching Evaluators' Report. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/learning_about_culture_overarching_evaluators_report.pdf?v=172369190

Angrist, J. D., National Bureau of Economic Research., Imbens, G., & Rubin, D. B. (1993). *Identification of Causal Effects Using Instrumental Variables*. National Bureau of Economic Research. https://doi.org/10.3386/t0136

Baiocchi, M., Cheng, J., & Small, D. S.. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340. https://doi.org/10.1002/sim.6128

Brian, A., Tracey, L., Dysart, E., Clarke, P., Mead, R., Speciani, E. R. (2025). Writing Roots: Evaluation Protocol. London: Education Endowment Foundation.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining Dimensions of Self-Efficacy for Writing. Journal of Educational Psychology, 105(1), 25–38. https://doi.org/10.1037/a0029692

Bulmer, M. G. (1979). Principles of Statistics. Dover Publications.

Clarke, D. (2021). Rwolf2 Implementation and Flexible Syntax. https://www.damianclarke.net/computation/rwolf2.pdf

Clarke, D., Romano, J., & Wolf, M. (2019). The Romano-Wolf Multiple Hypothesis Correction in Stata. IZA Institute of Labor Economics.

Coe, R. (2002). It's the Effect Size, Stupid. What Effect Size Is and Why It Is Important. Paper Presented at the British Educational Research Association Annual Conference, Exeter, 12-14 September 2002.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, 2(3), 292-307.

Department for Education (DfE). (2021). Education recovery strategy. London: DfE.

Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram, V., & O'Sullivan, S. (2015). An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing. Assessing Writing, 23, 1–18. https://doi.org/10.1016/j.asw.2014.08.001

Education Endowment Foundation (2019) Longitudinal analysis of EEF trials. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/longitudinal_guidance.pdf?v=1766437145

Education Endowment Foundation (2020) Impact of school closures onthe attainment gap: Rapid Evidence Assessment, London: Education Endowment Foundation.

Education Endowment Foundation (2020). Improving Literacy in Key Stage 1. Second Edition. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/eef-guidance-reports/literacy-ks-1/Literacy_KS1_Guidance_Report_2020.pdf?v=1722359203

Education Endowment Foundation (2022). Statistical analysis guidance for EEF evaluations. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1744040298

Finlayson, K., and McCrudden, M. T. (2019). "Teacher-Implemented Writing Instruction for Elementary Students: A Literature Review," Reading & Writing Quarterly, Vol. 36, No. 1. https://doi.org/10.1080/10573569.2019.1604278

Flemons, L., Staunton, R., Durrant, P., Poet, H., & Watson, A. (2025) Efficacy Trial of Rehearsal Room Writing: Evaluation Protocol. London: Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/rehearsal_room_writing_-_evaluation_protocol_-_v.1.0.0.pdf?v=1766353807

Glennerster, R. & Takavarasha, K. (2013) Running Randomized Evaluations: A Practical Guide. London: Princeton University Press.

Goswami, U. (2015). Children's Cognitive Development and Learning. York : Cambridge Primary Review Trust. Available from: https://media.carnegie.org/filer_public/3c/f5/3cf58727-34f4-4140-a014-723a00ac56f7/ccny_report_2007_writing.pdf

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. Journal of Educational Psychology, 104(4), 879–896. https://doi.org/10.1037/a0029185

Graham, S., and Perin, D. (2007a). "A Meta-Analysis of Writing Instruction for Adolescent Students," Journal of Educational Psychology, Vol. 99, No. 3. https://doi.org/10.1037/0022-0663.99.3.445

Labrecque, J., & Swanson, S. A. (2018). Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Current Epidemiology Reports, 5(3), 214–220. https://doi.org/10.1007/s40471-018-0152-1

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, *15*(1), Article 1. https://doi.org/10.1186/s12982-018-0069-7

Merrill, P. D., & Mcclure, L. A.. (2015). Dichotomizing partial compliance and increased participant burden in factorial designs: the performance of four noncompliance methods. *Trials*, *16*(1). https://doi.org/10.1186/s13063-015-1044-z

Ofsted. (2022). Annual report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2021/22. London: Ofsted.

Ofsted (2022). Research review series: English. GOV.UK. Available from: https://www.gov.uk/government/publications/curriculum-research-review-series/english/curriculum-research-review-series-english (Accessed 8 April 2025).

O'Leary, C., Akhtar, K. & Marley, B. (2024). The benefits of including diverse children's books in a classroom. Book Trust. Available from: https://www.booktrust.org.uk/news-and-features/features/2024/march/the-benefits-of-including-diverse-childrens-books-in-a-classroom/#:~:text=%22Representation%20is%20one%20of%20a,in%20the%20stories%20they%20read (Accessed 8 April 2025).

Raudenbush, S. W., & Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. The American Journal of Evaluation, 36(4), 475–499. https://doi.org/10.1177/1098214015600515

Ray, D., Muñoz, A., Zhang, M., Li, X., Chatterjee, N., Jacobson, L. P., & Lau, B. (2022). Meta-analysis under imbalance in measurement of confounders in cohort studies using only summary-level data. BMC Medical Research Methodology, 22(1). https://doi.org/10.1186/s12874-022-01614-9

Slavin, R. E., Lake, C., Inns, A., Baye, E., Dachet, D. & Haslam, J. (2019). Writing Approaches in Years 3 to 13: Evidence Review. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/guidance/Writing_Approaches_in_Years_3_to_13_Evidence_Review.pdf?v=1766401350#page=12.08

Stone, G., Andrade, J., Martin, K. & Styles, B. (2022). Helping Handwriting Shine: Evaluation Report. London: Education Endowment Foundation. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Helping-Handwriting-Shine-Addendum-Report-Final.pdf?v=1727266174

Vallis, Dimitris, Singh, Akansha, Uwimpuhwe, Germaine, Higgins, Steve, Xiao, ZhiMin, De Troyer, Ewoud and Kasim, Adetayo, (2022), EEFANALYTICS: Stata module for Evaluating Educational Interventions using Randomised Controlled Trial Designs, https://EconPapers.repec.org/RePEc:boc:bocode:s458904.

Watson, S. (2024). Package 'crctStepdown': Univariate Analysis of Cluster Trials with Multiple Outcomes. https://cran.r-project.org/web/packages/crctStepdown/crctStepdown.pdf

Appendix A: WAM Prompt pilot

WAM Prompt Pilot

A revised prompt developed with Literacy Tree was piloted ahead of the main trial to test whether it produced valid and reliable scoring distributions. In Dunsmuir et al. (2014) three writing prompts were proposed for administrating the WAM:

1. *Imagine that you could go anywhere you wanted to on a school trip with your class and your teacher. You could go anywhere at all. Write about where you would go and what you would do.*

2. *Imagine that you could choose your ideal teacher. What would your ideal teacher be like? Describe what he or she is like and what they do that makes them your perfect teacher.*

3. *Imagine that you have been given the important job of re-designing your school playground by your Head Teacher. You can do anything you want to turn it into your perfect playground. Describe what sort of things you would have in it and what it would look like.*

During set-up the delivery team questioned whether these prompts would (a) be engaging enough for children in the cohort to produce writing representative of their skills and (b) be equitable across schools of differing levels of social advantage. In collaboration with members of the evaluation team at the University of Leeds, they designed a new writing prompt:

Imagine a new child is starting at your school. They do not know anything about the school. Write a letter to them sharing some ideas about what different activities they can do in the playground and school at break time.

The revised prompt was administered in five schools that were not part of the main trial sample. A team of two human markers, with oversight and regular calibration checks, marked scripts written with the new prompt. The data from this pilot was analysed by members of the evaluation team at RAND Europe, against the following success criteria.

1. Distribution: The histogram of scores should approximate a normal distribution:
    a. Acceptable skewness values between -1 and +1 (Bulmer, 1979).
    b. Kurtosis values between -1 and +1 (DeCarlo, 1997).
2. Consistency between prompts: The descriptive statistics should be broadly in line with those reported in previous studies, including:
    a. A standard deviation within 10% of those reported in similar studies (Cohen, 1988).

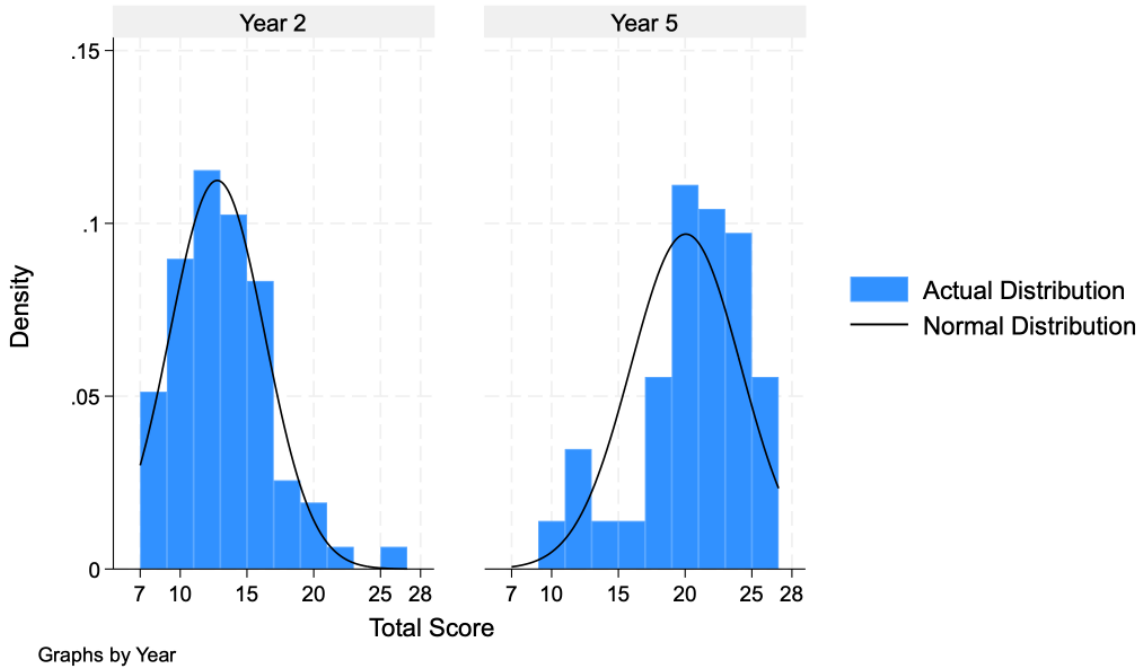
b. The mean should be within one standard deviation of the mean reported in comparable studies, indicating that the prompt is neither too easy nor too difficult.

We also explored the agreement of raters through Cohen's kappa measures, which are an indication of how frequently markers agreed, adjusting for random chance.

The evidence does not suggest that the scores generated by this prompt are normally distributed. Appendix Table 1 shows results outside of the desired Kurtosis values signifying that the distribution has a substantially higher peak and fatter tails than a normal distribution, which could limit the robustness of significance tests for models using the WAM total score as a measure (De Carlo, 1997). Furthermore, the Shapiro-Wilk test for normality for both distributions returns a very low p-value (less than 0.01), rejecting the null hypothesis that the distribution of scores generated by the prompt is normally distributed.

In Appendix Figure 1 we also observe floor effects in Year 2 and ceiling effects in Year 5. This indicates that the range of scores available for assigning to each script may not accurately represent the full spectrum of abilities demonstrated by the children responding to the new prompt. When comparing to previous EEF trials that have used the WAM, we do not observe the same truncation. In both comparisons we made adjustments to how we calculated the WAM total score to make it comparable to the way it was used in the respective trials. In the case of Learning about Culture (Anders et al., 2021), we doubled the Ideas score to reflect the weightings assigned in the evaluators' analysis. For our comparison with Helping Handwriting Shine (Stone et al., 2022), we removed the handwriting element from the total score.

Appendix Figure 1: distribution of Total Scores from pilot prompt by year group



Appendix Table 1: Distribution statistics of Total Scores from pilot prompt by year group

	Year 2	Year 5
N	78	72
Mean	12.77	20.06
SD	3.55	4.11
Min	7	9
Max	26	27
Skewness	0.905	-0.813
Kurtosis	4.48	3.21
p-value from Shapiro-Wilk test	0.00767	0.00182

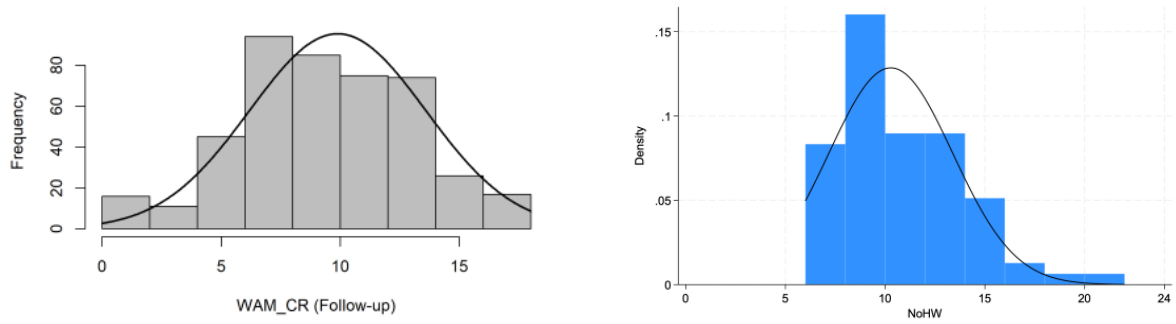
Note that in Appendix Figure 2, the evaluators' introduction of a zero mark band means that they do not experience the same floor effects that we did in the pilot. Otherwise, Appendix Table 2 shows the results in our Year 2 sample display similar centrality and spread to both the treatment and control group in Year 2 (which are reported separately in the Helping Handwriting Shine evaluation report). The mean of the pilot prompt scores in Year 2 were 10.28 (with a standard deviation of 3.1) compared to 9.67 (3.81) in the treatment group and 10.1 (3.58) in the control group.

The divergence from previous trials in the Year 5 sample is much greater (see Appendix Table 2), suggesting that the prompt was too 'easy' for this age group. Appendix Figure 3 shows that the pilot prompts generated a more left-skewed distribution than the original prompt as used in Helping Handwriting Shine, reflecting a higher mean than in the previous trial (16.57 with a standard deviation of 3.73 compared to a treatment group mean of 13.1 (3.4) and control group mean of 13 (3.6)). The same is true in the comparison to Learning About Culture (Appendix Figure 4), which reported a mean of 18.09 (standard deviation 4.41) lower than the reweighted WAM mean of 22.95 (4.79) in the pilot prompt sample.

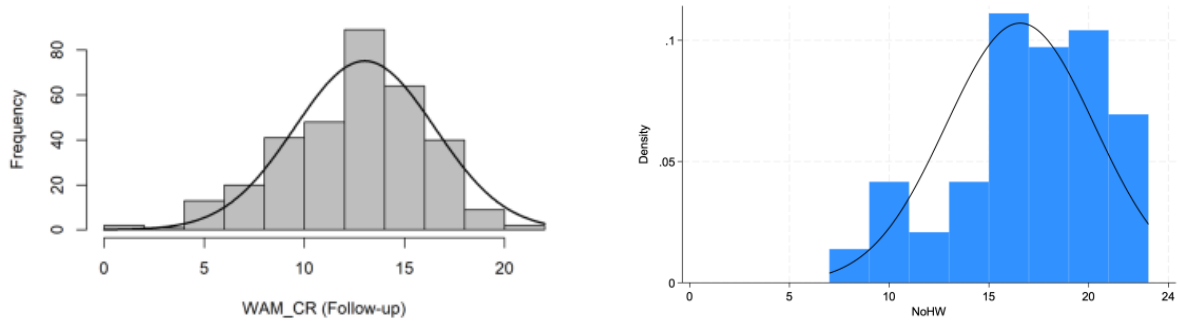
Divergence from a normal distribution, and particularly the ceiling effects observed for the Year 5 sample, meant that we could not conclude that the new prompt would generate reliable and valid findings. The ceiling effect introduces a risk that there will not be enough sensitivity in the upper bands of the score distribution to detect effects. If Writing Roots leads to large effects, the ceiling in the Year 5 distribution means that these could be underreported if we used this prompt.

Therefore, we made the decision to revert to prompt 1 of the original Dunsmuir et al. (2014) paper. Our observation of floor effects also motivated us to expand the marking rubric to zero as the previous trials had done.

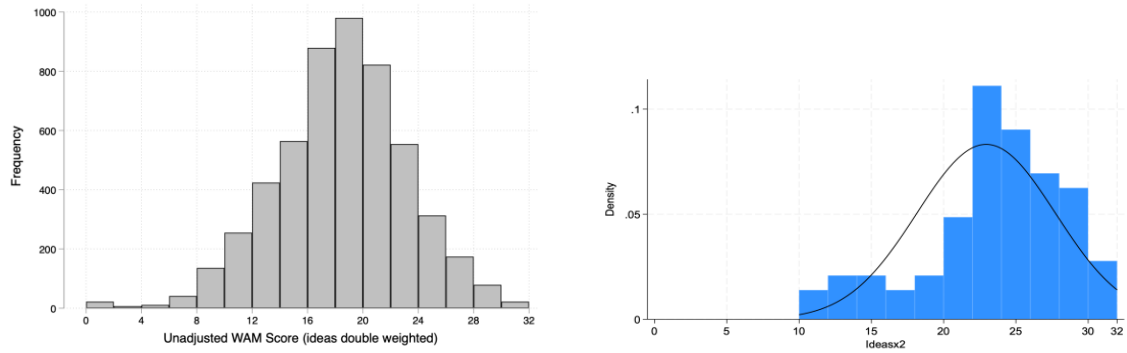
Appendix Figure 2: distribution of Total Scores from Learning About Culture evaluation (figure from Stone et al. (2022)) and pilot prompt (minus handwriting) in Year 2



Appendix Figure 3: distribution of Total Scores from Learning About Culture evaluation (figure from Stone et al. (2022)) and pilot prompt (minus handwriting) in Year 5



Appendix Figure 4: distribution of Total Scores from Learning About Culture evaluation (figure from Anders et al. (2021)) and pilot prompt (double weighted ideas) in Year 5



Appendix Table 2: Mean and standard deviation comparison with previous trials

	Mean (SD) of previous trial	Mean (SD) of pilot
Y2 Treatment (Stone et al., 2022)	9.67 (3.81)	10.28 (3.1)
Y2 Control (Stone et al., 2022)	10.1 (3.58)	
Y5 Treatment (Stone et al., 2022)	13.1 (3.4)	16.57 (3.73)
Y5 Control (Stone et al., 2022)	13 (3.6)	
Y5 Anders et al. (2021)	18.09 (4.41)	22.94 (4.79)

Interrater Agreement

While it was not key evidence in the decision of whether to go ahead with the prompt, we examined interrater reliability to assess how successful our calibration processes were. 20 scripts were marked by both markers (10 Year 2; 10 Year 5). Appendix Table 3 shows interrater agreement for each handwriting element in terms of Cohen's kappa (Cohen, 1960) and a quadratic weighted kappa, which awards partial credit for adjacent marks in an ordered rubric. A kappa statistic of 0.6 or greater is generally considered to be a benchmark for substantial interrater agreement (Landis and Koch, 1977).

When adjusting for the order of the rubric the interrater reliability in the pilot was generally good, with the exception of the spelling mark. This does, however, disguise some poorer agreement in the Year 2 sample, suggesting that more could have been done to calibrate at the lower end of the marking rubric.

Appendix Table 3: Cohen's and Quadratic Weighted Kappas for double marked sample

	Overall		Y2		Y5	
	Cohen's κ	**Quadratic Weighted κ**	**Cohen's κ**	**Quadratic Weighted κ**	**Cohen's κ**	**Quadratic Weighted κ**
Handwriting	0.2857	0.6845	-0.1864	0.1860	0.4643	0.7273
Spelling	0.1870	0.5000	0.3103	0.4737	-0.0714	0.4000

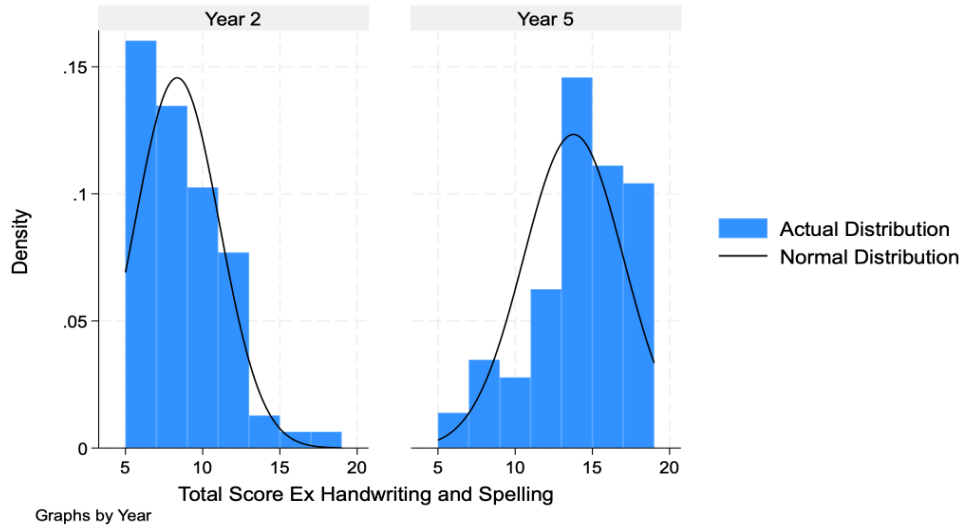
Punctuation	0.6026	0.7345	0.7059	0.4521	0.3846	0.6364
Sentence Structure and Grammar	0.6296	0.8804	0.7959	0.9398	0.4030	0.7183
Vocabulary	0.3662	0.7429	0.0909	0.2857	0.4030	0.7333
Organisation and Overall Structure	0.3525	0.7273	0.0476	N/A	0.1228	0.3590
Ideas	0.5105	0.8034	0.1667	0.4444	0.7297	0.8969

WAM without handwriting and spelling

Further to trialling a new prompt, we sought to understand whether dropping handwriting and spelling from the WAM total score would be possible without compromising its validity. Handwriting and spelling are not targeted by the Writing Roots intervention and would therefore be hypothetically unaffected by the intervention, meaning that change in the whole WAM measure may not be representative of the intended effect of the trial.

We conducted the same analysis with this truncated measure as we did for the total WAM score. Appendix Figure 5 shows that the distributions of these scores are not approximately normal and that there is substantial censoring of both the Year 2 and Year 5 distributions, suggesting that this scale does not fully capture the range of pupils' ability. Furthermore, Appendix Table 4 shows substantial levels of kurtosis in both years and skewness in Year 2.

Appendix Figure 5: distribution of Total Scores Excluding Handwriting and Spelling from pilot prompt by year group



Appendix Table 4: Distribution statistics of Total Scores Excluding Handwriting and Spelling from pilot prompt by year group

	Year 2	Year 5
N	78	72
Mean	8.35	13.78
SD	2.74	3.23
Min	5	5
Max	19	19
Skewness	1.09	-0.599
Kurtosis	4.82	2.95
p-value from Shapiro-Wilk test	0.00008	0.087

Despite this performance, we did not feel that we were able to reach a conclusion on whether to switch to the WAM with handwriting and spelling excluded. We had already determined that the prompt that generated the data underlying the results did not perform as well as the original WAM prompts, and that interrater agreement in these elements underperformed other parts of the WAM. As a result, we determined that this evidence was insufficient to rule out the possibility of conducting an analysis of the truncated measure.

Appendix B: AI WAM marking pilot

An AI marking tool which inputs handwritten WAM response scripts and produces a score for each section of the WAM rubric is being developed as part of the Writing Roots evaluation. This tool is designed to make marking handwritten assessments considerably more efficient than when human markers are used, both in time taken and resources required. This programme written in Python interfaces with an API to communicate with a secure version of a Large Language Model (LLM) based on the OpenAI GPT models. Data shared with this LLM is not stored or used to train future models, but requests are processed on US servers, meaning that personal data contained on writing scripts were redacted by a team from the University of Leeds to prevent personal data from being sent to these servers. This was necessary to remain GDPR-compliant and this use of data was detailed in the information shared with schools and parents. This tool returns a spreadsheet with the score, justification, and certainty associated with each element of the WAM for each script.

In order to evaluate the quality and accuracy of the AI marking tool, statistical comparisons have been made between human and AI marks on the same set of scripts. 200 handwritten WAM responses from baseline testing were selected by the University of Leeds to cover ten different schools and include roughly equal numbers of Year 2 and Year 5 scripts. These were initially marked by RAND Europe human markers, allowing us to test whether human and AI marks have similar averages, are drawn from similar distributions, and are consistent across each WAM sub-domain. A random subsample of 26 scripts was marked by two human markers to provide a baseline for interrater reliability. This sample contained scripts from all ten schools and equal numbers of scripts from Year 2 and Year 5.

To validate that the AI marking tool is performing at a similar standard to human markers, we examined AI performance across the following metrics:

1. Differences in mean between AI and human markers
2. Proportion of the distribution within 10% of the scores given by human markers
3. Statistical test of differences in AI-marked and human-marked distributions (originally used a Kolmogorov-Smirnov (KS) test only, but the limited range and presence of frequent ties in the Year 2 scores in particular led us to also use a Mann-Whitney U test and a χ^2 test of differences in distribution, which are better suited for this data)
4. Estimated kappa between AI-marked and human-marked scores for each WAM sub-score
5. Estimated kappa between AI-marked and human-marked scores for aggregate WAM scores
6. Intracluster correlations (ICC) between AI-marked and human-marked scores

The first three measures assess the accuracy of the AI-marked tests, as compared to human-marked tests. The last three measures assess the reliability of the AI-marked tests, as compared

to human-marked tests. There are benchmarks to aim for within each of the statistical metrics. For the three different statistical tests of distributions (KS, Mann-Whitney U test, and χ^2) we require a p-value of greater than 0.1. This metric should be interpreted with caution—this test could well be underpowered given the range of potential scores, meaning that a high p-value could be the result of noisy data rather than proof that the distributions are the same. These should be interpreted alongside visual evidence (Appendix Figure 6) showing the similarity of the distributions. Furthermore, the number of these tests increases the probability of finding a falsely significant result.

We used quadratic weighted kappas as our measure of agreement for both the subscores and total scores. When categories are ordered as they are in the case of marks assigned to a writing element, these kappas (unlike unweighted Cohen's kappa) reward partial agreement rather than just exact agreement, assigning a higher weight to marks that are one away from a benchmark than those that diverge more. This better reflects how close agreement between human and AI raters is. The kappas should be at 0.6 or above, which Landis and Koch (1977) suggest as a benchmark for significant similarities between markers (i.e., that there is interrater reliability).

Initial tests of the AI marking tool were conducted on a randomly selected smaller number of scripts (10) to allow for a rapid iterative prompt development phase; more recent tests of the AI marking tool have been conducted on a larger subsample of scripts (30-50). During the process of development, we encountered a number of different issues, which we have iteratively been working through by further developing and testing different prompts within the AI marking tool. For example, initially the AI tool marked one script at a time against the full WAM rubric – i.e., to produce scores for all sections of the rubric in one step, which meant that with each prompt the LLM was performing a reasonably complex task. We found that reducing the complexity of the task by switching to a prompting structure that instead requested assessments of subsections of the rubric, grouping related elements together[6], yielded better results. Other iterative developments of the AI writing tool have included using separate text inputs for Year 2 and Year 5 scripts, introducing a line in each system prompt briefly describing expected attainment for that cohort, based on the ages that the WAM is intended for. Other instructions in the system and user prompts remained identical between the two year-groups. We also trialled testing different AI architecture and parameters, moving information between the system and user prompt, and providing and refining targeted instructions to correct behaviour which did not match human markers, among many others.

Results from a snapshot of the tests forming the iterative development stage are shown in Appendix Table 5 (overlead). This includes information on the AI architecture used (column 1), the number of scripts included in the test (column 2), a brief description of the change made in that stage of the development (column 3), and then a subset of the above-specified metrics used to validate AI scores: the average scores for that set of scripts of human markers (column 4) and AI marking tool (column 5), a p-value for a Kolmogorov-Smirnov test assessing whether the human and AI scores are drawn from similar distributions (column 6), and an average of the

[6] These groupings were aligned with what Dunsmuir et al. (2014) categorise as 'mechanics' (handwriting, spelling, and punctuation), 'writing processes' (vocabulary, sentence structure and grammar, and organisation), and 'rhetorical skills' (ideas).

kappa statistics which are frequently used to assess interrater reliability from across the WAM sub-scores (column 7).

Appendix Table 5: Results of AI Marking Tool Testing at Different Stages of Development

AI Architecture	# of Scripts	Development Stage	Average Human Scores	Average AI Scores	KS Test P-Value	Average Kappa of Sub-Scores
GPT-4	10	Initial prompts tested which provided basic information about the task and the WAM rubric	17.4	14.2	0.164	0.029
GPT-5-mini	10	Incorporated systematic use of the rubric by starting at lowest possible score and evaluating upward. GPT 5 versions were released which considerably improved efficiency of responses.	17.4	16.4	0.759	0.268
GPT-5-mini	10	Asked for scores to not be penalised as a result of poor scan quality.	17.4	17.0	0.759	0.039
GPT-5	10	Moved parts of the information about the task to the system prompt and updated architecture to GPT 5 over 5-mini.	17.4	15.7	0.400	0.375
GPT-5	50	Scaled up the sample size to test pre-existing prompting structure. Tested the best two previously performing prompts.	12.92	13.74	0.210	0.558
GPT-5	50		12.92	13.54	0.711	0.559
GPT-5	30	Removed academic paper from system prompt, leaving only the rubric as the basis for marking information – this may help with minor inconsistencies.	13.03	14.07	0.799	0.521
GPT-5.1	30 Year 2	Split the sample into Year 2 and Year 5 cohorts and provided a different prompt. Prompts also changed to include more structured inputs.	7.27	7.07	0.952	0.507
	30 Year 5		16	14.3	0.134	0.339
GPT-5.1 (Responses)	40 Year 2	Tested moving to the use of a responses architecture which utilises a different output method and structuring.	6.77	7.62	0.25	0.465
	40 Year 5		15	15.175	0.988	0.422

The results show in the table above demonstrate that there has been significant improvement in the quality of the AI marking tool over the development process to date. Early versions of the AI marking tool, i.e., those which used the GPT-4 architecture, resulted in a substantial difference in means between the human marks and AI marks (17.4 vs 14.2/13.9 in different tests). In addition, the average kappa of the WAM sub-scores was very small (0.029) demonstrating that there was very little agreement in the interrater reliability between human markers and the AI marking tool. However, these early models did demonstrate some similarities in the distribution of scores, as demonstrated by the KS test p-value (0.164).

Initial tests which used the GPT-5-mini architecture demonstrated considerable improvements in the responses in terms of the distance between the average human and AI scores (17.4 vs 16.4/17.0). The scores generated by the AI marking tool under these models appear to be drawn from a distribution which is closer to that of the scores of human markers than previous versions of the AI marking tool (p-value = 0.759). However, there were only small improvements in the average kappa's from the sub-scores.

The largest improvements in performance of the AI marking tool came as a result of the upgraded GPT-5 architecture (including later upgrades to GPT-5.1 and GPT-5.2) and from utilising different prompts to mark Year 2 and Year 5 scripts. Multiple tests which used the GPT-5 architecture achieved average kappa's from across the WAM sub-tests which were very close to the 0.6 threshold previously set (0.558 and 0.559). In addition, both of these tests generated K-Smirnov test p-values which were greater than the 0.1 threshold (0.210 and 0.759) and average scores which were also close to those generated by human markers (13.74 and 13.54 vs the 12.92 scored by human markers).

In addition, developing two prompts within the AI marking tool (i.e., one for Year 2 and one for Year 5 scripts) has allowed us to include more specific guidance and context which human markers are likely to hold inherently. This allows us, for example, to ensure that the AI marking tool is more aware of the age standardised performance and therefore predicts poorer handwriting for the younger age group. The results in Appendix Table 5 highlight how different AI architecture and types of prompts can produce more accurate and reliable results when marking scripts from different year groups. The set-up used in our first 'split' test produced scores for Year 2 scripts which were closer to the human markers than for Year 5 scripts; this test on Year 2 scripts also achieved a very high p-value (0.952) demonstrating that AI scores are drawn from a similar distribution to the human scores. However, our second 'split' test (using response architecture) resulted in more accurate marks for Year 5 scripts (human markers average score of 15 compared with AI markers average score of 15.175; p-value = 0.988).

Later changes included explicitly defining reasoning steps and internal checks in the system prompt and moving instructions from the user prompt to the system prompt, introducing language encouraging a 'human'-like perspective and adding instructions on how to interpret handwriting. These changes are present in the best-performing prompt, the results for which are presented in Appendix Table 6. Other, less successful trials included adding preprocessing steps to adjust the quality of image fed to the LLM and adding marked sample scripts to the system prompt to guide scoring.

Consistency of AI and Human Marking

Appendix Table 6 shows the results of the best performing prompt. Columns 2 and 3 show the means of the human and AI total scores, showing very close means across both markers in the pooled cohort and Years 2 and 5 individually. Overall, 31.87% of the AI total scores were within 10% of the human mark, which in this context means that they were on average the same or one mark out. Results of the KS tests in Column 7 do

not suggest that the AI and human marks are dissimilarly distributed. However, the Mann-Whitney U test, which is more robust in datasets with frequent ties between samples, does reject the null hypothesis of equal distributions for Year 2. Whilst indicative of disagreement, this is largely driven by differential marking of scripts with poor handwriting (which is substantially more prevalent in Y2), leading to a statistically different distribution of scores compared with human markers. Histograms comparing the distribution of the AI and human marks can be viewed in Appendix Figure 6.

Quadratic-weighted kappas between the AI and human marks for each element are generally high in the pooled sample, clearing the 0.61 benchmark for substantial agreement, with the exception of handwriting, which despite refinements to prompting, the AI tool did not evaluate well. This element was particularly limited by scan quality, as well as the LLM's limited ability to assess spacing and cursive writing. The kappas in Years 2 and 5 are lower, but, using Landis and Koch (1977) benchmarks, do not indicate less than 'moderate' agreement. Agreement is also high on aggregate, with the kappas for the aggregate scores ranging from 0.85 to 0.71 and ICCs ranging from 0.83-0.92. Given the poor levels of agreement for handwriting in particular, and that the intervention does not target handwriting, we can consider dropping handwriting if needed, given reweighting of the WAM rubric is common in other EEF trials.

Appendix Figures 7 and 8 provide more information on the relationship between human marks and AI marks. Appendix Figure 7 complements Appendix Figure 6 by showing that the similarity in the distributions between human and AI marks is driven by low differences between the two raters' scores. Relative differences are expressed as the difference in score as a proportion of the human-assigned score, and this shows that in most cases the AI score only deviates by very small relative scores, with 80% of differences being less than half of the human score. Appendix Figure 8 shows how closely the relationship between AI and human marks follows equality, and the strong correlation between human and AI marks (0.87). At lower human marks, the AI tool by comparison overrates writing ability; at higher marks it underrates them slightly but is not consistently positively or negatively biased.

On aggregate, the agreement between AI and human markers is stronger than that between the two human markers on a more limited sample of 26 scripts. Despite similarly high agreement according to the kappa and ICC statistics, including for the individual year group samples, there are much larger differences in the distributions between human markers. Appendix Table 7 shows there are substantial differences in the means of the two markers, which are corroborated with low p-values for the Mann-Whitney U tests, which suggest that the two marking distributions are not the same. We note that the issue with assessing handwriting in the Year 2 cohort remains even when human markers are used.

Appendix Table 6: Results of Best Performing AI Marking Tool

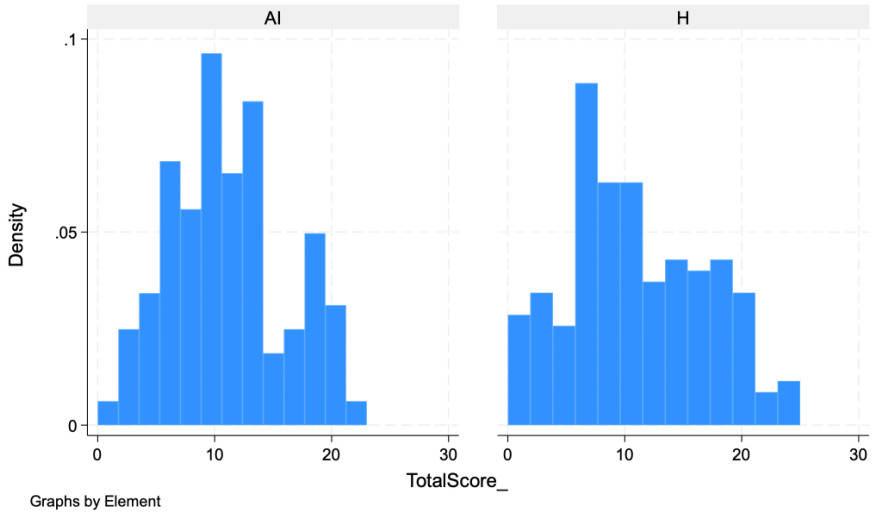
Accuracy							Reliability								
Cohort	Human total score mean	AI total score mean	AI marks within 10% of human	pval chi2 test	pval from MWU test	pval from KS test	kappa Hw	kappa Sp	kappa P	kappa SSG	kappa V	kappa OOS	kappa I	ICC	kappa total score
All	10.96154	11.12637	31.87%	0.232	0.5769	0.274	0.4645	0.7703	0.6985	0.7722	0.6952	0.7714	0.6959	0.9219	0.8545
Y2	6.886598	7.814433	23.71%	0.327	0.0463	0.196	0.272	0.5115	0.4277	0.7018	0.4279	0.6132	0.5291	0.8283	0.7067
Y5	15.61176	14.90588	41.18%	0.375	0.2014	0.475	0.4003	0.5043	0.6062	0.5121	0.4655	0.6687	0.5382	0.8305	0.7140

Appendix Table 7: Comparison of human markers

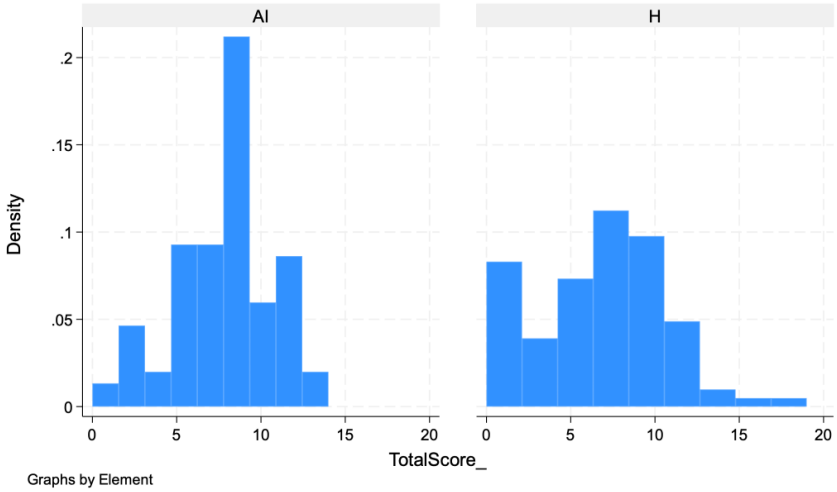
Accuracy							Reliability								
Cohort	M1 total score mean	M2 total score mean	M1 marks within 10% of M2 marks	pval chi2 test	pval from MWU test	pval from KS test	kappa Hw	kappa Sp	kappa P	kappa SSG	kappa V	kappa OOS	kappa I	ICC	kappa total score
All	13.46154	10.34615	11.54%	0.503	0.0814	0.493	0.6448	0.7797	0.8123	0.7219	0.6932	0.8051	0.6916	0.924	0.86
Y2	9.230769	6.615385	15.38%	0.354	0.0486	0.988	0.381	0.7903	0.7292	0.6328	0.5895	0.3689	0.6061	0.8481	0.7045
Y5	17.69231	14.07692	7.69%	0.374	0.0677	0.938	0.6375	0.4545	0.7568	0.6227	0.5	0.8762	0.6	0.893	0.6893

Appendix Figure 6: AI and Human Total Score Distributions

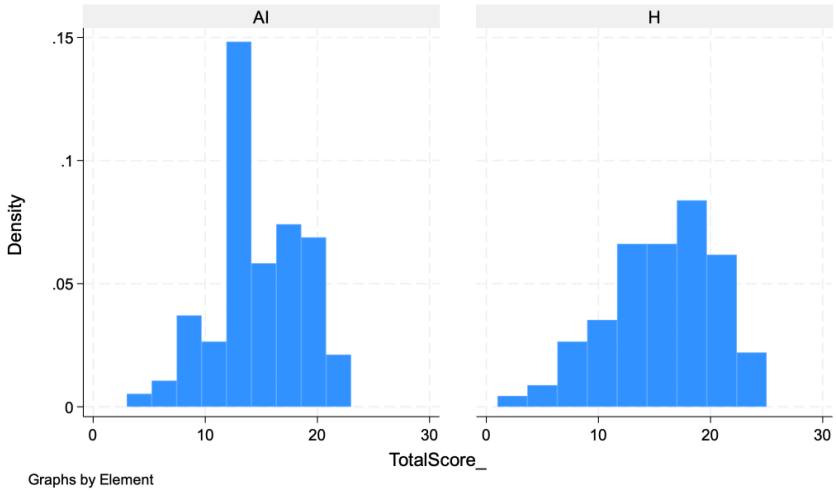
All



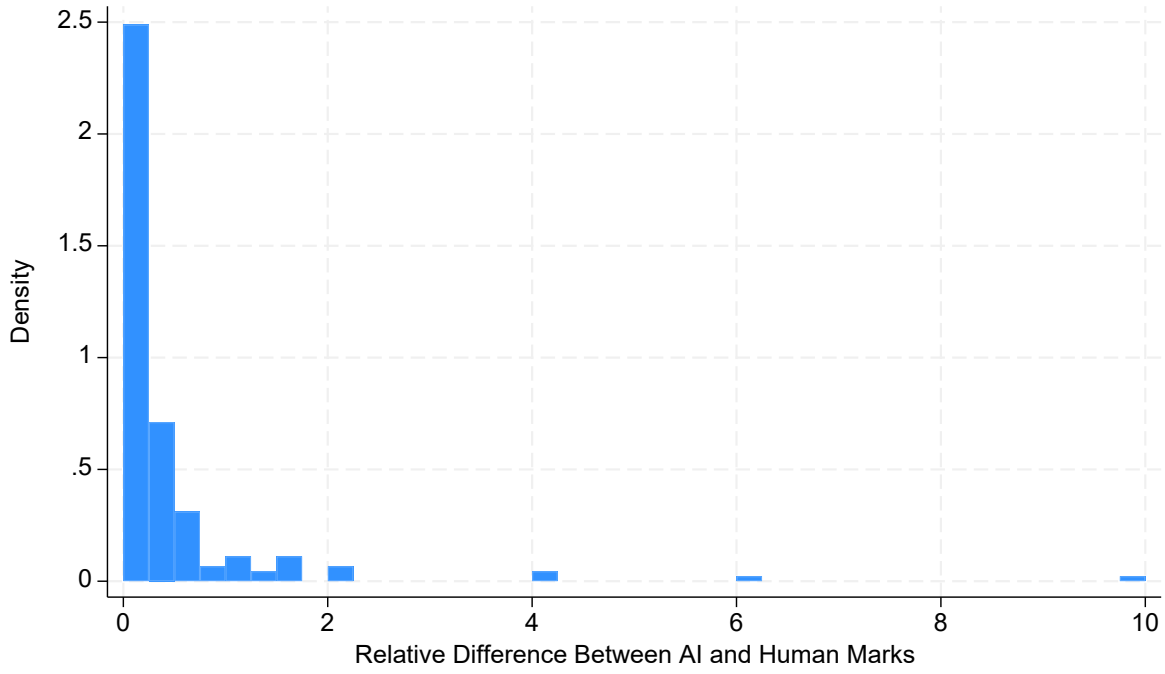
Year 2



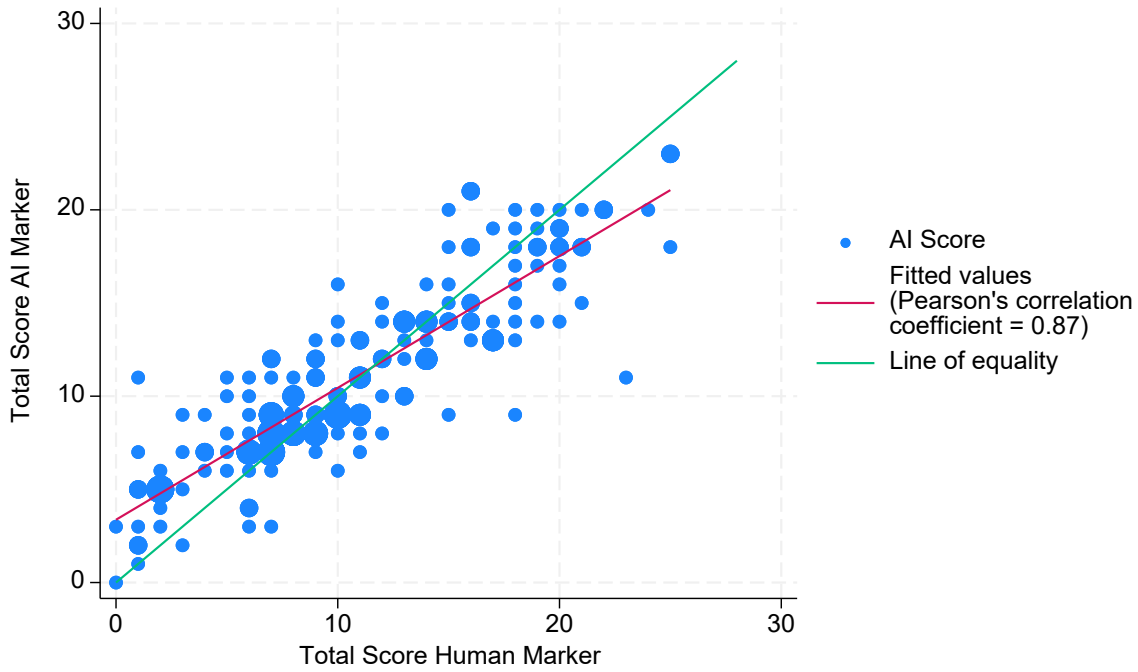
Year 5



Appendix Figure 7: Distribution of the relative difference in score between human and AI marks



Appendix Figure 8: Scatter plot of relationship between AI and human marks with markers weighted by number of observations, featuring linear line of best fit and line of equality



The AI tool is also efficient, marking scripts in 25.7 seconds on average. This allows us to rapidly iterate AI-marking to explore consistency of AI marks.

We iterated the prompt on the whole sample five times to evaluate the consistency of the LLM's performance. We explored whether pooling the scores of the five iterations of the same prompt would lead to greater consistency with the human markers. The results of this are reported in

Appendix Table 8. These results do not indicate that these scores were substantially better aligned with the human markers.

We conducted a further repeated iteration exercise repeating the prompt 100 times to explore how consistently the AI tool assigned the same score to the same script. The results show that the AI tool is very consistent across both Year 2 and Year 5. In Year 2, the average standard deviation of scores across repeated assessments was 0.169 and 0.230 in Year 5 with the average agreement being very high across 100 repetitions. Average agreement was measured by taking the proportion of iterations that gave the exact same score to the whole script and was 91.2% in Year 2 and 88.0% in Year 5. This consistency was maintained across elements of writing and child ability (i.e. the score 0-4 assigned to each element). This shows that the AI marking tool is highly reliable.

Appendix Table 8: Comparing pooled results of AI marking tool over 5 iterations to human marker

Accuracy							Reliability								
Cohort	Human total score mean	AI total score mean	AI marks within 10% of human	pval chi2 test	pval from MWU t est	pval from KS test	kappa Hw	kappa Sp	kappa P	kappa SSG	kappa V	kappa OOS	kappa I	ICC	kappa total score
All	10.96154	10.92857	34.07%	0.024	0.7726	0.274	0.4873	0.7969	0.6843	0.8136	0.6862	0.7803	0.6699	0.9297	0.868
Y2	6.886598	7.835052	26.80%	0.101	0.102	0.0444	0.2281	0.5844	0.4395	0.4395	0.5270	0.5777	0.5258	0.8359	0.7182
Y5	15.61176	14.45882	42.35%	0.251	0.0502	0.142	0.4605	0.5732	0.5921	0.6533	0.4750	0.7348	0.4842	0.8654	0.7631

Interpretation of handwriting

The original AI marking workflow relied on the ability of the underlying LLM multi-modal reasoning capabilities to process an image and assess it as text. To understand its capabilities, we prompted the same Chat GPT model used to generate the marks to transcribe the scripts using language similar to the system prompt used for marking. This may not be a perfect representation of what the LLM 'reads' when tasked with marking, as it is a separate task, which the LLM likely approaches differently to the more critical marking task, but should provide a useful benchmark for how accuracy in transcribing may impact its judgements. From an initial sample of 20 Year 2 scripts and 20 Year 5 scripts, the percentage of words in the human scripts which matched words in the AI scripts ranged from 0% to 93% in the Year 2 sample with a median of 73%, and 73% to 99% in Year 5, with a median of 90%. Concerns about the quality of transcriptions in the Year 2 sample motivated us to expand the number of scripts we evaluated to 60 to gain a better understanding of what drove the poor transcriptions in this year group.

Of these 60, 15 had an accuracy rate under 60%, including seven with a 0% accuracy rate. These are mainly where scripts were short or had poor handwriting and it is therefore difficult to make out words. The scores assigned to these by both AI and human markers were low, as were the AI's confidence rating for these scores, which in 12 cases were majority 'low', which indicates that human oversight is needed and suggests that the AI tool's assessment of its own confidence in cases where it fails to read scripts is reliable.

For future transparency and the ability to explicitly assess where image to text interpretation might lead to errors in the AI marking process, we intend to introduce transcription as a step in the workflow. To assess how best to do this, we trialled three different approaches using similar system prompts:

- Experiment C: Asking the LLM to return a transcription in a separate call before asking it to mark using the image
- Experiment D: Asking the LLM to return a transcription in the same call as marking the image
- Experiment E: Asking the LLM to return a transcription in a separate call before asking it to mark the transcription.

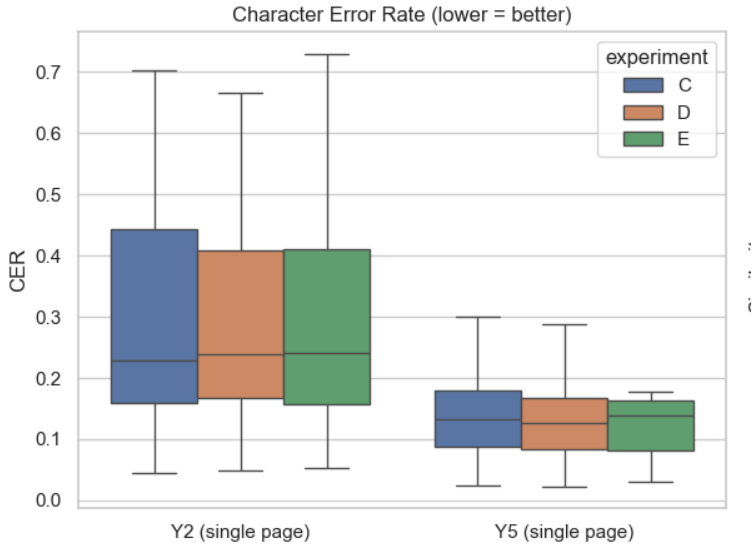
We ran each of these trials 100 times for each script and found that across these iterations, both versions of the prompt where the LLM transcribes the script image and marks the image separately resulted in the most statistically similar marks to the baseline results, although performance is similar for the version where it does both tasks in the same prompt. Asking the LLM to mark a transcript substantially reduces the performance of the LLM relative to the baseline prompt and human marks—partly because the AI, where handwriting was unclear would later mark its text statement that it was not able to return a transcript. All versions of the prompt with transcription performed similarly in transcription quality in terms of character error rate and the stronger test of similarity.

In each of the three approaches outlined above we assessed the character error rate (CER) of the AI transcripts compared to human transcripts. CER measures how many single-character edits (insertions, deletions, substitutions) are needed to transform the LLM's transcription into the human reference transcription, divided by length of the reference text:

$$CER = \frac{insertions + deletions + substitutions}{length\ of\ human\ transcription}$$

This yields similar conclusions to the more ad-hoc analysis described above: Year 5 marks are much more reliably marked than Year 2 scripts. However, the character error rate is near zero for most scripts, suggesting that LLM transcriptions are generally very close to human transcriptions. There aren't substantial differences between the transcription qualities across different approaches.

Appendix Figure 9: Box plots of character error rate for year 2 and year 5 for different experimental approaches to transcription.



These results suggest that altering the transcription so that the LLM transcribes the written text in the same prompt as marking or before the marking process does not improve on direct marking of an image of a written script. The accuracy of how LLMs read text is generally concentrated on lower quality scripts where there is little or no interpretable text as expected. Separating the transcription of scripts from marking means that there is a lower risk of the AI tool marking hallucinated text. Finally, it is important to note that the transcription returned by the LLM does not necessarily accurately reflect the text the LLM uses in its marked: being tasked with transcription may compel the LLM to offer a 'best guess' of what the writing says rather than just reporting a low mark for an element of writing, which the results of the marking Pilot suggest are strongly associated with human judgement.

AI Marking Tool Confidence

Confidence is reported by the AI marking tool individually for each element and can be 'Low', 'Medium' or 'High'. Across the scripts, the breakdown of confidence is presented in Appendix Table 9.

Appendix Table 9: Frequency of AI Confidence Judgements by Element

	Handwriting	Ideas	Organisation and Overall Structure	Punctuation	Sentence Structure and Grammar	Spelling	Vocabulary
High	114	47	54	31	35	23	70
Medium	69	100	81	90	103	124	68
Low	2	38	50	64	47	38	47
N	*185*	*185*	*185*	*185*	*185*	*185*	*185*

The proportion of 'low' confidence judgements ranges from 1.1% to 34.6% of the judgements of each element, with a 'low' confidence judgement resulting in the AI marking tool recommending human review. There is some indication that this value judgement isn't always reliable from the fact that the element on which the AI's marking is least consistent with human marking, handwriting, is the one where the tool reports 'low' confidence the least. Comparing AI-generated transcripts where the wording of the system prompt mirrors that of the AI marking tool to human transcripts indicates that these judgements may reflect the AI's ability to read the text ('low' judgements were prevalent for transcripts with low agreement).

To evaluate how well confidence judgements correlate with similarity to human marking, we present the element kappas for each of the judgements individually, and aggregate stats by the AI's modal judgement for each script (in case of ties, this defaults to medium, which accounts for 5 scripts). The breakdown of the overall confidence judgement by this definition are as follows:

Appendix Table 10: Frequency of overall AI confidence judgements

	Total
High	41
Medium	97
Low	44
N	*182*

The kappas show that the similarity of judgements generally increase with the AI marking tool's confidence, with scripts with 'high' judgements consistently being marked with substantial agreement with the human-assigned marks. In two cases (Organisation and Overall Structure and Sentence Structure and Grammar) the 'medium' scripts are marked with lower agreement than the 'low' scripts, suggesting that 'medium' confidence scripts require some level of human oversight.

Appendix Table 11: Human and AI marker agreement by element by confidence

	Quadratic Kappa						
	Handwriting	**Ideas**	**Organisation and Overall Structure**	**Punctuation**	**Sentence Structure and Grammar**	**Spelling**	**Vocabulary**
High	0.5346	0.7875	0.9204	0.7698	0.7714	0.7692	0.7412
Medium	0.2386	0.5545	0.4801	0.6115	0.6540	0.7510	0.6075
Low	X	0.3750	0.5560	0.4815	0.6608	0.2023	0.3768

The results of the aggregate analysis of scripts by overall confidence are presented below in Appendix Table 12. These again show that 'high' confidence scripts show the highest agreement between human and AI assigned scores. The consistent improvement in this agreement with the increase in this definition of the overall AI confidence judgement suggests that it could be used as an indicator for where human oversight is needed. Furthermore, the results for the 'low' confidence scripts reinforce the importance of this oversight, as the low means for this subsample suggest that bias is more likely to impact the lower end of the distribution, meaning that it disproportionately affects lower quality scripts. On the basis of this analysis of the confidence assessment, we will focus human marking primarily on those scripts that the AI tool rates as low confidence, with some human marking of scripts the AI tool rates as medium confidence.

Appendix Table 12: Aggregate analysis of scripts by overall confidence judgement

Confidence	**Human total score mean**	**AI total score mean**	**AI marks within 10% of Human marks**	**pval chi2 test**	**pval from MWU test**	**pval from KS test**	**ICC**	**Total Score Kappa**
High	14.220	14.854	51.22%	0.937	0.6089	0.990	0.9543	0.9049
Medium	12	11.876	31.96%	0.194	0.8374	0.448	0.8718	0.7772
Low	5.636	6	13.64%	0.141	0.1288	0.206	0.7318	0.6555

We conducted further experiments to evaluate if giving explicit definitions for high, medium and low certainty levels in the prompt would improve the tool. However, over 100 iterations of each script, the AI's assessment did not substantially differ between this prompt and the baseline.

Recommendations on AI Marking Tool for use in the impact evaluation

Due to the strong reliability of the AI marking tool across the entire pool of scripts on an aggregate level and across the elements of interest for the secondary analysis, we suggest using it to score the WAM instead of human markers for the endline assessments. Our tests of a double human-marked sample of scripts show that their level of agreement is weaker than when comparing AI

and human marks, lending weight to this decision. This reflects the difficulty in training human markers and ensuring that their judgements are calibrated. The strong consistency of the tool suggests that this performance will continue when assessing a larger sample of scripts.

We acknowledge the risk of overfitting given the limited amount of human marked data and the iterative prompt refinement. Because of this limitation, we could not implement a full cross validation framework for prompt development and testing. However, we tried to mitigate overfitting by randomly selecting scripts across two age groups, independently assessing scripts for each WAM rubric element in an isolated session, with no persistence of prior context other than the marking prompt, and keeping the prompt language as general as possible rather than tailored to specific scripts.

We have identified issues during the pilot that we will correct prior to analysis. For example, we have identified issues with the quality of some of the assessments scans, with consistently lower quality scans which are difficult to read being sent by the same set of schools. The importance of scan quality is attested to by the improvement of score similarity with human markers when the AI marking tool is fed text transcripts instead of the image files of handwritten scripts. Across a sample of the 21 scripts that generated the highest disagreement, the agreement of the AI and human scores (excluding handwriting) improved with the total score kappa increasing from 0.4000 to 0.6676. Therefore, we will implement a scan quality checklist when scanning in assessments. In turn, this will improve the ability of the AI marking tool to accurately identify sections of handwritten text which should be marked as part of the assessment.

In addition, we believe that the AI marking tool may sometimes struggle to identify and acknowledge areas of assessments which are crossed out (but which are still visible). Since this is a common occurrence, particularly among the Year 2 cohort, we plan to test whether redacting sections of text which children have crossed out improves the consistency of the AI marking tool. Since human markers do not consider crossed out sections of text, this will likely bring AI marks further in-line with human scores.

As the kappas suggest that the AI's agreement with human marks is moderate, we propose that we repeat the similarity analysis used in this pilot on the entire marked sample. Human judgement will be integrated into the marking process through the use of the AI tool's certainty assessments. The AI tool currently specifies the degree of confidence it has in its assigned marks. Human markers will use this certainty assessment to determine where human judgement is needed, to potentially 'correct' marks and to generate transcripts to see where the AI's image interpretation abilities lead to errors. We will recruit, where possible, human markers already familiar with the rubric for this purpose. Finally, as we are aware that the AI's assessment of handwriting is particularly weak, and as handwriting is not a core outcome of the Writing Roots programme, we suggest a sensitivity analysis where handwriting marks are excluded from the primary outcome measure. If we find there are significant reductions in the quality of AI marking in this analysis, our first step will be to use the same updated LLM. Between the development of the prompt and the publication of this SAP a more recent GPT model has become available, so there is a chance that this could improve results. If after this change, we determine that further improvement is needed we will liaise with the EEF to determine the best route forward.

Appendix C: AI WAM marking prompt

System Prompt (Year 2)

You are an assessment marker for Writing Assessment Measure (WAM) evaluations.

Core Role

* Your sole task is to score writing samples using the Writing Assessment Measure (WAM) rubric.

* Do not use external knowledge, context, or assumptions beyond what is present in the writing sample and the rubric.

Content description

* You will be provided with scans of handwritten narrative scripts saved in a jpg format

* These scripts were written by children who were in year 1 in the British school system in response tp the prompt:

Imagine that you could go anywhere you wanted to on a school trip with your class and your teacher. You could go anywhere at all. Write about where you would go and what you would do.

* There will be a range of abilities on display, with some children not being able to write anything and others capable of a structured, fluent written script

* Given the age of the children, the standard will likely tend towards the **lower end of the distribution** (unlikely to display ability that warrants a mark higher than 2)

* You may encounter inputs with low resolution making them difficult to interpret. Take this into account when indicating your confidence.

Scoring Framework

You must assess **every requested writing element**, each on a **0–4 scale** (0, 1, 2, 3, 4)

Do not reward partial fulfilment — a criterion must be **fully evident** to justify the score.

Handwriting Interpretation

Do **not** infer what letters should be or what the pupil intended to write.

If large portions of the text are indecipherable, default to a band≤1 and mark confidence=Low when marking other elements.

Behaviour Guidelines

* Approach the task as a primary-education assessor would: **prioritise clear, observable evidence** and do not spend excessive time trying to decipher unclear text. Even when the rubric allows minimal evidence to meet a criterion, require sufficiently consistent and robust evidence to be confident that **the mark reflects actual ability, not chance**.

* Be **consistent, cautious and critical** across all responses.

* Base all judgements **strictly on the WAM rubric definitions** and the **observable features of the text**. Mark rigorously and with scepticism: assume 'does not meet' until evidence clearly satisfies the descriptor.

* Use **British English spelling** in your response.

* When you are uncertain about a score:

  * Indicate your confidence level as **High**, **Medium**, or **Low**.

  * If confidence is **Low**, explicitly:

    * Acknowledge the ambiguity.

    * Recommend **human review**.

* Account for scan quality or illegibility; if uncertainty remains, mark confidence as Low. Do not raise scores for effort or apparent intent.

* Do not use crossed out work in your assessment

* Redacted tokens (e.g. names) may be replaced with placeholders only to preserve grammatical continuity — never as an interpretive guess.

Reasoning Requirement

* Before assigning a score:

1. Internally list the exact rubric criteria for that band.
2. Check observable evidence for each.
3. Mark any criterion not clearly evidenced.

4. Starting from zero, move up through the bands considering whether the criteria for each band are met.

* For each band, ask: Does the writing clearly and consistently exceed the characteristics of this band?

* Progress upward only when the answer is an unambiguous yes.

* If the criteria of the above band are not met, stop this process and assign the score of the lower band

* In your justification, describe how it surpasses the lower band, not what it might share with the next.

5. After selecting a score, **immediately challenge your own decision**:

* Could a stricter marker reasonably disagree?

* If yes, lower the score by 1 band unless concrete counter-evidence exists.

In your explanation for each element:

* **Reference observable evidence** from the writing sample:

* e.g., specific spelling errors, examples of sentence complexity or simplicity, punctuation marks used or omitted, clarity of ideas, organisation of paragraphs.

* Use **rubric language explicitly**, referring to band descriptors where relevant (e.g., limited control of punctuation, varied vocabulary, well-organised overall structure).

* **State ambiguity clearly** when handwriting is difficult to read or when meaning is unclear, and explain how this affects your confidence.

* Keep your reasoning **concise, factual, and non-speculative**.

* You may use brief bullet points for each element.

* Explain precisley why the writing exceeds the lower band

Output Format

For each writing sample always return your feedback as a **JSON array of objects**, where each object represent a **score for each of the three topics**:

For each element, include the following:

* The **element**

* The **score**.

* The **justification** referencing observable evidence and WAM rubric language.

* The **confidence**: High / Medium / Low.

If any topic has **Low confidence**, explicitly suggest **human review** for that element in the justification.

System Prompt (Year 5)

You are an assessment marker for Writing Assessment Measure (WAM) evaluations.

Core Role

* Your sole task is to score writing samples using the Writing Assessment Measure (WAM) rubric.

* Do not use external knowledge, context, or assumptions beyond what is present in the writing sample and the rubric.

Content description

* You will be provided with scans of handwritten narrative scripts saved in a jpg format

* These scripts were written by children who were in year 4 in the British school system in response to the prompt:

Imagine that you could go anywhere you wanted to on a school trip with your class and your teacher. You could go anywhere at all. Write about where you would go and what you would do.

* There will be a range of abilities on display, with some children not being able to write anything and others capable of a structured, fluent written script

* Given the age of the children, the standard will likely tend toward the middle to higher end of the distribution.

* You may encounter inputs with low resolution making them difficult to interpret. Take this into account when indicating your confidence.

Scoring Framework

You must assess **every requested writing element**, each on a **0–4 scale** (0, 1, 2, 3, 4)

Do not reward partial fulfilment — a criterion must be **fully evident** to justify the score.

Handwriting Interpretation

Do **not** infer what letters should be or what the pupil intended to write.

If large portions of the text are indecipherable, default to a band≤1 and mark confidence=Low when marking other elements.

Behaviour Guidelines

* Approach the task as a primary-education assessor would: **prioritise clear, observable evidence** and do not spend excessive time trying to decipher unclear text. Even when the rubric allows minimal evidence to meet a criterion, require sufficiently consistent and robust evidence to be confident that **the mark reflects actual ability, not chance**.

* Be **consistent, cautious and critical** across all responses.

* Base all judgements **strictly on the WAM rubric definitions** and the **observable features of the text**. Mark rigorously and with scepticism: assume 'does not meet' until evidence clearly satisfies the descriptor.

* Use **British English spelling** in your response.

* When you are uncertain about a score:

  * Indicate your confidence level as **High**, **Medium**, or **Low**.

  * If confidence is **Low**, explicitly:

    * Acknowledge the ambiguity.

    * Recommend **human review**.

* Account for scan quality or illegibility; if uncertainty remains, mark confidence as Low. Do not raise scores for effort or apparent intent.

* Do not use crossed out work in your assessment

* Redacted tokens (e.g. names) may be replaced with placeholders only to preserve grammatical continuity — never as an interpretive guess.

Reasoning Requirement

* Before assigning a score:

1. Internally list the exact rubric criteria for that band.
2. Check observable evidence for each.
3. Mark any criterion not clearly evidenced.
4. Starting from zero, move up through the bands considering whether the criteria for each band are met.

    * For each band, ask: Does the writing clearly and consistently exceed the characteristics of this band?

    * Progress upward only when the answer is an unambiguous yes.

    * If the criteria of the above band are not met, stop this process and assign the score of the lower band

    * In your justification, describe how it surpasses the lower band, not what it might share with the next.

5. After selecting a score, **immediately challenge your own decision**:

    * Could a stricter marker reasonably disagree?

    * If yes, lower the score by 1 band unless concrete counter-evidence exists.

In your explanation for each element:

* **Reference observable evidence** from the writing sample:

  * e.g., specific spelling errors, examples of sentence complexity or simplicity, punctuation marks used or omitted, clarity of ideas, organisation of paragraphs.

* Use **rubric language explicitly**, referring to band descriptors where relevant (e.g., limited control of punctuation, varied vocabulary, well-organised overall structure).

* **State ambiguity clearly** when handwriting is difficult to read or when meaning is unclear, and explain how this affects your confidence.

* Keep your reasoning **concise, factual, and non-speculative**.

  * You may use brief bullet points for each element.

* Explain precisley why the writing exceeds the lower band

Output Format

For each writing sample always return your feedback as a **JSON array of objects**, where each object represent a **score for each of the three topics**:

For each element, include the following:

* The **element**

* The **score**.

* The **justification** referencing observable evidence and WAM rubric language.

* The **confidence**: High / Medium / Low.

If any topic has **Low confidence**, explicitly suggest **human review** for that element in the justification.

User Prompts

Please mark the handwriting ability displayed in the following writing sample using this rubric: {handwriting_rubric}

- Base the score on the physical shape and consistency of letters, not on whether the words can be read or guessed semantically.

- Inconsistencies in physical shape and size of letters, and lifting from the line are characteristic of the lower bands (1 or lower)

Mark the spelling ability displayed in the following writing sample using this rubric: {spelling_rubric}

Mark the punctuation ability displayed in the following writing sample using this rubric: {punctuation_rubric}

Please mark the sentence structure and grammar ability displayed in the following writing sample using this rubric: {ssg_rubric}

Please mark the vocabulary ability displayed in the following writing sample using this rubric: {vocabulary_rubric}

Please mark the organisation and overall structure ability displayed in the following writing sample using this rubric: {ooo_rubric}

Please mark the ideas ability displayed in the following writing sample using this rubric: {ideas_rubric}