Effectiveness Trial of White Rose Maths Reception Jigsaw (2025 – 26), a Two-Arm Cluster Randomised Trial
Statistical Analysis Plan

Evaluator (institution): National Foundation for Educational Research (NFER)
Principal investigator(s): Pippa Lord

PROJECT TITLE	Effectiveness Trial of White Rose Maths Reception Jigsaw (2025 – 26), a Two-Arm Cluster Randomised Trial
DEVELOPER (INSTITUTION)	White Rose Education (WRE)
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Pippa Lord
SAP AUTHOR(S)	Chris Morton, Aarti Sahasranaman, Pippa Lord
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	4 – 5 years, Reception (Early Years)
NUMBER OF SCHOOLS	330 (target); 304 (randomised)
NUMBER OF PUPILS	6,600 (target); 6,494 (randomised)
PRIMARY OUTCOME MEASURE AND SOURCE	Early numeracy skills (Early Years Toolbox Early Numeracy App, Howard et al., 2022)
SECONDARY OUTCOME MEASURE AND SOURCE	Early mathematical skills (Early Years Foundation Stage Profile maths ELGs from National Pupil Database (NPD))

SAP version history

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

Table of contents

Introduction

The White Rose Maths (WRM) Reception Jigsaw programme, developed and delivered by White Rose Education (WRE), is a Professional Development (PD) programme that aims to improve early maths teaching quality and outcomes in schools.

The efficacy of Reception Jigsaw was previously evaluated using a cluster randomised controlled trial (RCT) commissioned by the Education Endowment Foundation (EEF) (Tang *et al.*, 2024). The trial found that children in Reception Jigsaw (RJ) schools made, on average, an additional one month's progress in maths compared to children in control schools. Although this finding had a high security rating, it was not statistically significant, i.e., the evaluator could not conclude that this represented a real effect. Children whose Reception class teacher complied with the programme requirements (i.e., attended at least nine out of the 10 training sessions) made an additional one month's progress. This finding was statistically significant and highlighted that Reception Jigsaw had a positive impact when compliance to the programme was achieved. Children eligible for Free School Meals (FSM) in intervention schools made no additional progress in maths compared to FSM children in control schools.

This effectiveness trial of the Reception Jigsaw programme has a cluster randomised design similar to the efficacy trial. Schools were randomly assigned to receive either the Reception Jigsaw programme ('intervention' schools) or continue with their usual practice ('control' schools). There are, however, two important design differences between the efficacy and effectiveness trials. The efficacy trial was not powered to detect an impact on FSM pupils, so one of the objectives behind increasing the size of this effectiveness trial was to improve the power in the FSM subgroup, within WRE's maximum delivery capacity. Another key difference between the two trials is the primary outcome measure. NFER used the New PUMA (Hodder's Progress in Understanding Mathematics Assessment) as the primary outcome measure in the efficacy trial. However, taking into account the issues with New PUMA and the feedback from WRE and schools (Tang *et al.*, 2024), we have chosen the Early Years Toolbox (EYT) Early Numeracy assessment as the primary outcome measure. The secondary outcome is wider mathematical skills of Reception pupils measured by the Early Years Foundation Stage Profile (EYFSP), specifically the two Early Learning Goals (ELGs) for maths. To minimise the burden and cost of data collection, only a sample of up to 20 children across the Reception year group will be assessed for impact in each school.

Intervention schools receive access to the Reception Jigsaw programme during the 2025 – 26 academic year. The programme is targeted to Reception practitioners (teachers and teaching assistants), and attendance of the Maths Lead or Early Years Lead and other senior leaders is encouraged by WRE, where this is possible. Teachers receiving professional development then lead maths teaching and provision for pupils throughout the Reception year. The training received by the schools comprises five modules, each focusing on a specific aspect of early years maths and delivered during the course of the academic year. Each module uses the same cycle of development, comprising a two-hour PD training session, followed by setting a gap task which supports teachers to use their learning to develop early years maths in their school, followed by a half-day coaching and mentoring visit for personalised and tailored support. Attendance at the five PD sessions and the five coaching sessions informs the compliance measure for this trial (see Compliance section). All training sessions, coaching and mentoring visits happen face-to-face with a designated WRE trainer visiting the school.

Research questions

Primary research question

- RQ1. What is the impact of Reception Jigsaw on early numeracy (measured using the Early Years Toolbox) at the end of Reception?

Secondary research questions

- RQ2. What is the impact of Reception Jigsaw on disadvantaged pupils' early numeracy (measured using Early Years Toolbox) at the end of Reception?
- RQ3. What is the impact of Reception Jigsaw on pupils' score on the Early Years Foundation Stage Profile (EYFSP) maths Early Learning Goals (ELGs)?
- RQ4: How does teachers' training/coaching attendance/compliance impact pupils' early numeracy (primary outcome)?

Design overview

Trial design, including number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		Government Office Region, to support the delivery of the Reception Jigsaw programme
Primary outcome	variable	Early numeracy (maths attainment)
	measure (instrument, scale, source)	Early Years Toolbox Early Numeracy App, 0 – 120, (Howard et al., 2022)
Secondary outcome(s)	variable(s)	Early mathematical skills
	measure(s) (instrument, scale, source)	EYFSP, binary variable for whether pupil meets the two mathematics ELGs (1) or not (0), National Pupil Database.
Baseline for primary and secondary outcome	variable	Maths attainment
	measure (instrument, scale, source)	Teacher assessment based on observation (Emerging Numeracy checklist), 10–60, bespoke instrument

This is an effectiveness trial to assess the impact of the WRM Reception Jigsaw programme on Reception pupils. The evaluation uses a two-arm cluster randomised controlled trial design. Participating schools will be randomly allocated in a 1:1 ratio into two arms, stratified by Government Office Region (GOR) to reduce the chance of WRE trainers having too many or too few schools delivering the programme relative to their capacity. Schools in the 'intervention' arm

receive the WRM Reception Jigsaw CPD programme during the 2025-26 academic year. Schools in the 'control' arm continue with their usual practice during the 2025-26 academic year.

The complete pupil selection process is described in the study protocol, but is briefly summarised here. Pupils at participating schools who are in Reception for the 2025/26 academic year will be eligible to participate in the trial. There are no specific eligibility restrictions, but only a sample of Reception pupils from control and intervention classes will be assessed for impact, as described below.

We first sampled up to 24 Reception pupils at random across all participating classes in the school for baseline Emerging Numeracy (EN) checklist completion (see 'baseline measures' below) in summer 2025. The EN checklist was completed by the sampled pupil's class teacher. At the 304 schools that went on to be randomised (so completed at least some EN checklists), 366 pupils (out of a total of 6860 sampled pupils) did not have a completed EN checklist so did not progress further through the selection process. Although we did not collect data on the reasons why EN checklists were left uncompleted, possible reasons include lack of staff time to complete all the checklists and exclusion of pupils that were perceived as unsuitable for the intervention (despite there being no specific pupil eligibility restrictions, as noted above). This shouldn't affect the internal validity trial results, as EN checklists were all returned before randomisation. However, it might slightly affect the external validity of results if pupils with very low prior ability had EN checklists completed with a disproportionately low frequency compared to their peers.

At this point, 234 schools (70% of those randomised) had completed the EN checklist for more than 20 pupils (i.e. 21-24). However, only 20 of these Reception pupils will sit the EYT test at endpoint to maximise cost-efficiency relative to information yield and reduce testing burden on schools. These 20 pupils will be selected at random[1] from amongst the 24 pupils that were baselined and who are present on the day of endpoint testing. Any extra pupils above the 20 can effectively be used as substitutes for those with missing endpoint data (e.g. absent on the testing day), increasing the analysed sample size.

Outcome measures

Baseline measures

The baseline measure corresponding to the primary outcome and EYFSP secondary outcome is the Emerging Numeracy checklist (see Appendix B of trial protocol for the full checklist). The EN checklist was developed for use in the Reception Jigsaw efficacy trial (Tang *et al.*, 2024) as data from the Reception Baseline Assessment (RBA) is not available to researchers via the NPD. This bespoke checklist was developed in consultation with colleagues in the Centre for Assessment at NFER and early years specialists at White Rose Maths, to gauge numeracy skills in children in the early years without placing undue burden on teachers and pupils. It consists of 20 tasks based on the Early Years Outcomes and ELGs. The checklist is completed by teachers about their pupils - ideally after they have completed the RBA or after two weeks of the pupil attending school. It takes 5–10 minutes to complete per pupil. A Cronbach's alpha of 0.95 suggests it is a reliable measure. It had a correlation of 0.59 with the outcome measure in the Reception Jigsaw

[1] The project statistician will randomly select 20 pupils for endpoint testing and the remaining 1-4 pupils will act as substitutes for any of the 20 that are absent on the day of testing. A priority order for the substitutes to be used will be determined, also at random. The random selection is therefore performed in advance, but exactly which pupils are tested is determined by who is present on the day.

efficacy trial and so we consider it a more cost-effective way to improve statistical power, compared to using a commercial test at baseline (even though this might have slightly higher correlation). As mentioned previously, we only require the EN checklist be completed for a sample of pupils across the Reception year group, to reduce both teacher completion burden at baseline and follow-up cost of test administrators.

Primary outcome

NFER used the New PUMA as the primary outcome measure in the efficacy trial of Reception Jigsaw. However, New PUMA was not chosen as the primary outcome measure in this effectiveness trial for two important reasons. First, following changes to the Reception curriculum in 2021, it was recognised that alignment between New PUMA and the revised curriculum was not strong. Second, feedback from WRE and some schools suggested that some aspects of this measure were not in line with expectations of Reception-age pupils or with teaching approaches followed in Reception Jigsaw. Following an extensive exploration of other measures, we concluded that the Early Years Toolbox Numeracy assessment (EYT) is the most suitable measure for this effectiveness trial. The EYT is more aligned than the New PUMA with the current Reception curriculum and so is also more aligned with the RJ programme, which was developed for that curriculum. The EYT has been widely implemented across numerous early childhood settings across multiple states in Australia (Howard and Melhuish, 2017); in England, the EYT was the subject of a pilot study (Dawson *et al.*, 2020) co-funded by the Department for Education and was used as an outcome measure in the Study of Early Education and Development (SEED) study (Melhuish and Gardiner, 2020).

The EYT measures key areas of early development including numeracy, language, self-regulation, social emotional development and executive function.[2] Each measure is a game-like assessment that is brief (5–10 minutes to complete) and will likely sustain the attention and interest of young pupils. Each measure is downloaded as an app for iPads. The Numbers measure covers children's emerging numeracy skills, including numerical language, spatial and measurement concepts, counting, matching digits and quantities, completing number lines, ordinality, subitising, patterning, numerical word problems and equations. The EYT Numeracy assessment has demonstrated construct validity, concurrent validity with established measures and high test-retest reliability (Howard *et al.*, 2022).

Secondary outcome

We will use an aggregation of the two mathematics ELGs from the Early Years Foundation Stage Profile (EYFSP) as a secondary outcome measure. If pupils meet the expected level in both the number and numerical patterns ELGs, this secondary outcome will be a 1 (ELGs met); if pupils do not meet the expected level in one or both ELGs it will be 0 (ELGs not met). If one or both ELGs has missing data the overall outcome will be missing and the pupils will not be included in this analysis. The NPD variables 'FSP_MAT_E11' and 'FSP_MAT_E12' will be used to determine whether each ELG has been met. This will provide another, though wider (i.e. less proximal), measure of mathematics attainment (expected/emerging level of development).

Sample size calculations overview

All sample size calculations were performed using the PowerUpR package (Bulus *et al.*, 2021) in the R statistical software (The R Foundation, 2025), using the function 'mdes.cra2'. This function

[2] https://www.eytoolbox.com.au/download.html

computes the MDES for a two-level (pupil, school) cluster-randomised trial. As there are more than one class at some schools (see 'Additional analysis' for more discussion), arguably a three-level sample size calculation would be more appropriate. However, there are additional parameters in the three-level calculation (e.g. the class-level ICC). Estimates for these parameters would likely be inaccurate, as there are relatively few previous educational trials that have used three-level models. the two-level approach was also informed by operational considerations, particularly the decision to sample pupils across rather than within classes.

We required a trial design that could detect a relatively small intervention effect, given the effect size of 0.08 seen for the primary outcome in the RJ efficacy trial (Tang *et al.*, 2024). Because the primary outcome measure may be more appropriate for Reception pupils (see 'Primary outcome') the effect size observed could increase relative to the efficacy trial. Additionally, the impact of RJ in the efficacy trial may have been affected by Covid-related disruption, although the trial itself did not find evidence of this, based on data collected in the IPE. However, we also note that effect sizes are typically smaller in effectiveness trials than at the efficacy stage. The net result of these three changes is uncertain: we decided on target MDES of 0.137, which was limited by the maximum number of schools WRE could deliver to.

Original trial design

Originally, the evaluation design was for 14 Reception pupils per school to be selected for the trial, with 10 completing the EYT endpoint assessment. The choice of including no more than 10 pupils per school was made for cost-efficiency reasons - as a maximum of 10 EYT assessments can be completed per test administrator per day - and to reduce the testing burden on schools. Including more schools in the trial was considered a more cost-efficient way to ensure the trial was adequately powered than including more than 10 pupils per school. At this stage, the recruitment target was 400 schools, which is higher than most EEF-funded trials, but was considered achievable from a recruitment perspective. This number of schools allowed for an MDES of 0.137, after accounting for attrition and specifying further parameters as given for 'Current trial design' below[3].

During the recruitment phase, it became increasingly clear that recruiting 400 schools before the autumn 2025/26 term would not be possible. This was in part due to a pause in active recruitment for 6 weeks, due to internal restructuring within WRE and confirmation of the continuation of RJ in its current form. By July 2025, it was deemed possible to recruit approximately 330 schools, rather than the 400 intended. This prompted discussions around whether the trial could be redesigned to achieve the originally intended MDES of 0.137 with only 330 schools. The only way to achieve an MDES of 0.137 was by increasing the number of pupils assessed using the EYT, requiring two endpoint testing visits per school. This was agreed with EEF, and pursuing this strategy led to the current trial design, which is described below.

Current trial design

For the current trial design, 24 pupils will have a baseline EN checklist completed, of which 20 will be assessed using the EYT at endpoint. The number of schools recruited at the protocol stage has been reduced to 330, to reflect the revised recruitment target after the recruitment pause described above. We estimated ICC, R_1^2 and R_2^2 values based on the endpoint data from the RJ

[3] The exception is the anticipated school-level attrition rate, which was reduced from 10% to 5% between the original trial design and the current trial design. This was due to the evaluators reconsidering the probable attrition rate, rather than new information becoming available about the attrition rate.

efficacy trial (Tang *et al.*, 2024)[4]. Although we are proposing a different outcome measure for this effectiveness evaluation, many other factors are the same, including the baseline measure, intervention and pupil year group, so the efficacy trial was considered the best available source for these parameters.

The specified parameters result in an MDES of 0.132 before factoring in attrition, as shown in the table below. In the FSM subgroup, the MDES was 0.170, assuming an average of 3.5 FSM pupils per Reception year group. This average was based on 17.5% of the 20 reception pupils analysed being eligible for FSM (see DfE statistics). We expected 5% school-level attrition and 15% pupil-level attrition, which would increase the MDES at the analysis stage to 0.137 (0.181 in the FSM subgroup).

Of the 359 signed Memorandums of Understanding shared by WRE, 304 schools were randomised, somewhat lower than the target of 330. Where schools gave a reason for withdrawing before randomisation, the most common reasons were work commitments and lack of time (N=13) and that the cost of the intervention was prohibitive (N=7). Due to this reduction in schools relative to the target, the MDES after factoring in anticipated attrition increased to 0.142 (0.186 in the FSM subgroup). The average cluster size at randomisation was 21.36: at schools with over 20 pupils, the cluster size will fall to a maximum of 20 pupils at endpoint, due to a maximum of 20 pupils being tested per school (see 'Design Overview'). This reduction is considered to be part of the anticipated 15% pupil-level attrition.

668 pupils were confirmed by schools as being eligible for FSM and 1,571 were confirmed as not eligible. However, there was a lot of missing data in this variable (N=4,255), as schools often did not know the FSM status of their pupils at the start of the Reception term. We therefore expect the number of FSM pupils to increase considerably by the analysis stage: FSM status on the NPD will be more complete and as it is collected later more FSM-eligible pupils will have been identified. We decided that multiplying the number of pupils randomised by 0.175 (the national FSM proportion) would give a more accurate estimate than using FSM data provided by schools, so this is what was done at the randomisation stage in the table below.

		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)		0.132	0.137	0.170	0.181	0.136	0.142	0.174	0.186
Proportion of variance explained	level 1 - pupil (R_1^2)	0.396		0.396		0.396		0.396	
	level 2 – school (R_2^2)	0.147		0.147		0.147		0.147	

[4] These numbers were calculated directly from the trial data so cannot be found in the report itself. The ICCs used in the report were conditional on model covariates, whereas those used here are unconditional.

		Protocol				Randomisation			
		OVERALL		FSM		OVERALL		FSM	
Intracluster correlations (ICCs)	level 2 (class)	0.183		0.189		0.183		0.189	
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		20		3.5		21.36		3.74	
Number of schools	intervention	165	157	165	157	152	144	152	144
	control	165	157	165	157	152	144	152	144
	total	330	314	330	314	304	288	304	288
Number of pupils	intervention	3300	2669	577.5	468	3225	2467	564	455
	control	3300	2669	577.5	468	3269	2501	572	462
	total	6600	5338	1155	936	6494	4968	1136	917

	School level	Pupil level
Expected attrition (%)	5%	15%

Randomisation

Schools that completed the EN checklists for one or more sampled pupils were eligible for randomisation. Randomisation occurred at the school level, with schools randomised into the intervention and control arms in a 1:1 ratio. Randomisation was carried out by an NFER statistician using R Code, which was quality assured by another statistician and stored for reproducibility and transparency (code provided in Appendix A). Neither statistician was blinded to group allocation. The allocation data was then passed over to NFER's Operations team, who liaised with schools and WRE. Randomisation was stratified by Government Office Region (GOR) to reduce the chance of some trainers needing to deliver the Reception Jigsaw training to more schools than their capacity allows.

Analysis

All analysis will be conducted in accordance with the EEF statistical analysis guidance. Aside from the compliance analysis, an Intention-To-Treat (ITT) approach will be followed throughout. This means pupils will be analysed according to their intervention or control group assignment, regardless of their degree of participation in the intervention. Analysis will be conducted on complete cases only: pupils with missing values for any variables used in an analysis model will be excluded from modelling. The missing data analysis will investigate the sensitivity of results

to this choice. The analyst will not be blinded to the intervention assignment for any part of the analysis.

Analysis will be performed using the R software (The R Foundation, 2025), with all mixed effects models analysed using the R package 'lme4' (Bates *et al.*, 2015).

Primary outcome analysis

The purpose of the primary analysis is to estimate the impact of RJ on early numeracy - as measured by the EYT - at the end of Reception (RQ1). A two-level (pupil and school) linear mixed effects model will be used for this analysis:

$$EYT_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 EN_{ij} + \beta_3 London_j + \cdots + \beta_{10} North\_East_j + b_j + \epsilon_{ij}$$

In this model:

EYT_{ij} =total EYT score of pupil i in school j, measured at endpoint;

β_0 =intercept term;

$intervention_j$ =indicator for whether school j was randomised to the intervention (1) or control (0);

β_2 =average impact of each additional point of the baseline EN score on the EYT score;

EN_{ij} = total EN score of pupil i in school j, measured at baseline;

$London_j, \dots, North\_East_j$ =eight indicator variables that take the value 1 if school j is in the given GOR and 0 otherwise (the 9th GOR, Yorkshire and the Humber[5], is used as the reference level);

$\beta_3, \dots, \beta_{10}$ = eight coefficients representing the average difference in EYT score between a pupil in the given GOR and the reference GOR (Yorkshire and the Humber);

b_j = school-level error term (random intercept);

ϵ_{ij} =pupil-level residual error term.

The GOR indicators are included as covariates, as they were used to stratify the randomisation. The main estimate of interest is β_1, which represents the average impact of Reception Jigsaw on EYT score, with a 95% confidence interval for this estimate calculated using the profile likelihood. The estimate and confidence interval will be converted to a standardised effect size, as described in 'Estimation of effect sizes' below.

Secondary outcome analysis

The secondary analysis will investigate the impact of RJ on whether pupils reach the expected standard in both maths ELGs in the EYFSP. To do this, a two-level (pupil, school) logistic regression will be run:

$$log(\frac{P_{ij}}{1 - P_{ij}}) = \beta_0 + \beta_1 intervention_j + \beta_2 EN_{ij} + \beta_3 London_j + \cdots + \beta_{10} North\_East_j + b_j$$

[5] The choice of reference level is arbitrary and does not affect the intervention estimate.

In this equation log is the natural (to the base e) logarithm function and P_{ij} is the probability that pupil i in school j meets the expected level in both ELGs. The remaining terms are defined above for the primary analysis.

The sample of pupils will be the same as the primary analysis, aside from differences due to missing outcome data. To minimise the burden of baseline EN checklist completion for schools, a maximum of 24 pupils will be analysed per school, despite ELG data being available on the NPD for all Reception pupils. Having baseline data for the sampled pupils improves the power for this outcome, compared to including all Reception pupils (but without baselines included in modelling).

Subgroup analyses

The NPD variable 'EVERFSM_6_P' from the school census will be used to identify whether pupils are eligible for FSM for all impact analyses, as recommended by the EEF statistical analysis guidance. This variable measures whether a pupil has been eligible for FSM at any point in the last six years. However, FSM data is only recorded from the Reception year onwards, so there are no prior years available to include here, and we therefore expect the variable to effectively measure FSM status in Reception only. This means 'EVERFSM_6_P' is likely to be a less reliable proxy for disadvantage than when the variable is used in later year groups.

The effect of the RJ programme amongst disadvantaged pupils (RQ2) will be explored by repeating the primary analysis model, restricted to the subset of pupils that are eligible for FSM.

$$EYT_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 EN_{ij} + \beta_3 London_j + \cdots + \beta_{10} North\_East_j + b_j + \epsilon_{ij}$$

We will also investigate the differential effect of the programme for FSM pupils relative to non-FSM pupils using the model:

$$EYT_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 EN_{ij} + \beta_3 FSM_{ij} + \beta_4 FSM_{ij} * intervention_j + \beta_5 London_j + \cdots + \beta_{12} North\_East_j + b_j + \epsilon_{ij}$$

The coefficient for the interaction term β_4 estimates the differential effect of the RJ programme for FSM pupils compared to non-FSM pupils: a positive value suggests the programme is more effective for FSM pupils, while a negative value suggests it is less effective. As discussed above, the FSM indicator measured in Reception is likely to underestimate the number of FSM pupils compared to later years, which may reduce the size of any difference observed from this model.

Further analysis

At the point of randomisation, there were 511 Reception classes included in the trial, which is an average of 1.68 per school. In many schools, there is therefore a three-level data structure, with pupils nested within classes and classes nested within schools.

We considered whether the primary analysis should be a three-level model, or alternatively, whether a separate 3-level model should be run as a sensitivity analysis. The three-level model would generally be expected to increase power compared to a two-level model, with the size of the increase depending on the size of the class-level ICC (Demack, 2019). This does, however, require that the class identifier is correctly specified. As we collected class data before the start of the Reception autumn term, it is possible that there were some changes to class allocations after that point. There is also a small possibility of model estimation issues for the three-level model due to schools with one class or classes with few pupils. However, simulation results

provided by Bruyndonckx et al. (2018) provide some reassurance that this will not be a problem, although the authors only examined two-level models. Considering these factors together, we decided that the primary analysis will remain two-level but a three-level model will be run as a sensitivity analysis. The sensitivity analysis will use the following three-level (pupil, class, school) linear model:

$$EYT_{ijk} = \beta_0 + \beta_1 intervention_k + \beta_2 EN_{ijk} + \beta_3 London_k + \cdots + \beta_{10} North\_East_k + a_{jk} + b_k + \epsilon_{ijk}$$

In this equation i indexes pupils within a class, j indexes classes within a school and k indexes schools. The model is similar to the primary analysis, except that now a_{jk} is a class-level random intercept, in addition to the school-level random intercept b_k.

Imbalance at baseline

To assess imbalance at baseline school- and pupil-level characteristics a table will be produced describing the following characteristics of the control and intervention groups after randomisation.

School level:

- Proportion of pupils eligible for FSM (ever eligible in the last six years)
- Proportion of pupils with special educational needs (SEN)
- Total headcount (full- or part-time attendance) of pupils in the Reception year
- Whether the school is urban or rural
- School type (academy, maintained or independent)
- Most recent overall Ofsted rating[6]
- Proportion of pupils meeting the expected standard in their KS2 maths exam
- Government Office Region

Pupil level:

- FSM eligibility
- SEN status
- Proportion of half-days absent in the autumn and spring terms[7]
- Gender

[6] Due to reforms to Ofsted inspections, from November 2025 an overall Ofsted rating is no longer given, only ratings across multiple domains. As most schools won't yet have a rating under the new system, we will use their most recent overall rating under the old system.

[7] The proportion will be calculated as $\frac{Numberofhalf-dayabsences}{Numberofhalf-dayspossible}$ using the NPD variables 'OverallAbsence_2Term' and 'SessionsPossible_2Term' both here and for the missing data analysis below.

- Baseline EN score

Categorical variables will be described in terms of counts and proportions, while means and standard deviations will be given for continuous variables. School-level variables will be obtained via publicly available data releases from the Department of Education, such as Get Information About Schools[8]. Time-variant school-level variables will be measured for 2025/26. Pupil-level variables will be obtained from the 2025/26 spring census within the NPD (FSM, SEN, absence rate, gender) or collected directly from schools (EN score). The difference between the baseline EN scores in the intervention and control groups will be estimated using a two-level (pupil, school) linear model and expressed as a standardised effect size.

Wherever national data is available we will also describe the same school and pupil-level characteristics for the 'target' population: the wider population for which this trial seeks to draw conclusions. At a school level this is all primary schools in England (LA-maintained, academies and free schools) and at a pupil level it is all Reception pupils in England. By comparing characteristics in the trial sample with those in the target population we can assess how representative the sample is, which may help infer the external validity of trial results.

Missing data

The number and proportion of missing cases in the primary analysis will be reported. If this is less than five percent then the potential for bias in a complete case analysis will be considered minimal and no further missing data analysis will take place, following EEF's statistical analysis guidance. Otherwise the analysis will proceed as described below.

We expect that there will only be missing data in the EYT outcome and not the EN baseline, as a completed EN checklist was a condition of randomisation. Where EYT scores are missing we will report the reasons why, using top-level absence codes collected by NFER test administrators (e.g. "pupil absent on the day of testing").

A mixed effects logistic regression with two levels (pupil and school) will be run in which the outcome will be the logit probability of the EYT outcome being missing. All other variables from the maths primary analysis model will be included as covariates, together with the following additional variables that may be associated with missingness:

- FSM eligibility
- Gender
- SEN status
- Proportion of half-days absent in the autumn and spring terms
- Number of pupils at the school per full-time teacher
- Proportion of pupils eligible for FSM

These variables will be measured in the 2025/26 academic year. The additional variables which demonstrate an association with missingness in EYT score, as indicated by a p-value below 0.05, will be included as covariates in the primary analysis model. If re-running the primary analysis with these extra covariates included alters the substantive interpretation of the intervention

[8] https://get-information-schools.service.gov.uk/Downloads

effect, then the EYT outcome may be 'missing at random' conditional on the inclusion of those covariates. This would then need to be discussed in the report.

The EEF statistical guidance indicates that sensitivity analysis using multiple imputation is appropriate when the baseline measure (here EN score) is missing, but this is not the case here, as only the EYT outcome will be missing.

Compliance

Compliance will be defined by attendance at the training sessions and coaching visits, which will be collected using an Excel training log. For modelling purposes, each pupil's compliance will be determined by the number of PD training sessions attended by their class teacher (0-5) AND the number of coaching visits attended by their class teacher and/or the key person at their school[9] (0-5). The five PD training sessions and five coaching and mentoring visits are all assumed to be of equal value and will be summed, so each pupil's compliance value will be a number between 0 and 10. The key person is identified by the school as someone (e.g. Maths Lead, senior leader) who will consistently attend all training sessions to provide continuity if the Reception class teacher(s) is unable to attend all training sessions and coaching visits. WRE do not consider it an essential part of the intervention that every class teacher attends every RJ coaching visit: the role of the key person is important in ensuring the learning from the training is incorporated in teaching practice when the pupil's class teacher is unable to attend.

The causal effect of the number of PD training and coaching sessions attended will be explored using instrumental variable modelling, estimated using two-stage least squares, in which the individual observations are pupils (not teachers). This causal effect is estimated for pupils who would have at least one PD training and coaching sessions attended if randomised to the intervention group (similar to the Complier Average Causal Effect for a binary compliance variable). The causal effect relies on the exclusion restriction—that the intervention influences outcomes only through compliance as described above—while recognising that other pathways cannot be completely ruled out. In particular, dichotomising the number of sessions at a given threshold to create a binary measure assumes there is no effect of the intervention beneath that threshold. We instead assume the relationship between the number of sessions and EYT score to be linear, which we believe to be a more realistic assumption.

For the first stage, the compliance variable will be regressed on the intervention indicator and primary analysis covariates. This first stage linear regression will be:

$$compliance_{ij} = \beta_0 + \beta_1 intervention_j + \beta_2 EN_{ij} + \beta_3 London_j + \cdots + \beta_{10} North\_East_j + \epsilon_{ij}$$

Here $compliance_{ij}$ is the number of PD training and coaching sessions attended. This will take the value zero for control pupils, as the trial design ensures control pupils cannot receive the intervention ('one-sided' compliance). For the second stage, EYT scores are regressed on each pupil's predicted compliance value $\hat{compliance}_{ij}$ obtained from the first stage, in the following linear regression:

$$EYT_{ij} = \beta_0 + \beta_1 \hat{compliance}_{ij} + \beta_2 EN_{ij} + \beta_3 London_j + \cdots + \beta_{10} North\_East_j + \epsilon_{ij}$$

[9] That is, for each coaching visit to be counted it is sufficient for the pupil's class teacher to attend, or for the school's key person to attend or both.

The coefficient for predicted compliance β_1 in this second stage can be interpreted as the average change in EYT score per additional PD training or coaching session attended. This coefficient will be presented both on its raw scale (EYT score change per session attended) and multiplied by ten to estimate the impact of 'complete' compliance compared to none.

Results from both regression stages will be reported. All instrumental variable analyses will be performed using the R package 'ivreg' (Fox *et al.*, 2021). These models do not include school-level random effects, so instead cluster-robust standard errors will be calculated using the R package 'sandwich' (Zeileis, 2006; Zeileis, Köll and Graham, 2020). The correlation between the intervention indicator and the number of PD training and coaching sessions, as well as the F-test for the intervention indicator from the first stage of the two-stage regressions, will be reported.

Intra-cluster correlations (ICCs)

Two-level (pupil, school) ICCs will be calculated as the proportion of variance attributable to school-level variation:

$$ICC = \frac{\sigma_S^2}{\sigma_P^2 + \sigma_S^2}$$

Here $\boldsymbol{\sigma_s^2}$ and $\boldsymbol{\sigma_p^2}$ are the school-level and pupil-level variance, which can be extracted directly from a mixed effects regression fitted by the 'lme4' package. The following ICCs will be calculated:

- For the EYT primary outcome, conditional on primary analysis covariates
- For the EYT primary outcome, unconditional (no covariates)
- For the EN checklist baseline, unconditional
- For the ELG secondary outcome, conditional on secondary analysis covariates
- For the ELG secondary outcome, unconditional

Additionally, the class-level ICC ($\boldsymbol{ICC_C}$) and school-level ICC ($\boldsymbol{ICC_S}$) will be calculated from the sensitivity analysis which uses a three-level model (see 'Additional analysis').

$$ICC_C = \frac{\sigma_C^2}{\sigma_P^2 + \sigma_C^2 + \sigma_S^2} \quad ICC_S = \frac{\sigma_S^2}{\sigma_P^2 + \sigma_C^2 + \sigma_S^2}$$

σ_C^2 is the class-level variation. These figures may help inform power calculation parameters for future trials that use three-level models. In this case the ICCs will be calculated three times:

- For the EYT primary outcome, conditional on sensitivity analysis covariates
- For the EYT primary outcome, unconditional (no covariates)
- For the EN checklist baseline, unconditional

Effect size calculation

Impact estimates from the linear models described above will be presented as an effect size, as described by Hedges (2007):

$$ES = \frac{\hat{\beta}}{\sqrt{\sigma_P^2 + \sigma_S^2}}$$

$\hat{\beta}$ is the coefficient for the binary predictor of interest, typically the intervention indicator, which will be extracted from a conditional model (including any covariates). σ_s^2 and σ_p^2 are the school-level and pupil-level variation from a model with the same outcome, but only the binary treatment covariate included. To obtain a 95% confidence interval for the effect size, a confidence interval for $\hat{\beta}$ will first be calculated. The end points of this confidence interval will then be divided by the denominator in the above effect size formula.

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67, pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Bruyndonckx, R., Hens, N. and Aerts, M. (2018) 'Simulation-based evaluation of the linear-mixed model in the presence of an increasing proportion of singletons', *Biometrical Journal*, 60(1), pp. 49–65. Available at: https://doi.org/10.1002/bimj.201700025.

Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: power analysis tools for multilevel randomized experiments'. Available at: https://cran.r-project.org/web/packages/PowerUpR/index.html (Accessed: 12 August 2025).

Demack, S. (2019) *Does the classroom level matter in the design of educational trials? A theoretical & empirical review*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Does_the_classroom_level_matter.pdf?v=1727426126 (Accessed: 7 October 2024).

Fox, J., Kleiber, C., Zeileis, A. and Kuschnig, N. (2021) 'ivreg: instrumental-variables regression by "2SLS", "2SM", or "2SMM", with diagnostics'. Available at: https://cran.r-project.org/web/packages/ivreg/index.html (Accessed: 27 January 2025).

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

Howard, S., Neilsen-Hewett, C., de Rosnay, M., Melhuish, E. and Buckley-Walker, K. (2022) 'Validity, reliability and viability of pre-school educators' use of early years toolbox early numeracy', *Australasian Journal of Early Childhood*, 47(2), pp. 92–106. Available at: https://doi.org/10.1177/18369391211061188.

Melhuish, E. and Gardiner, J. (2020) *Study of early education and development (SEED): impact study on early education use and child outcomes up to age five years*. Available at: https://assets.publishing.service.gov.uk/media/5e4e5c10e90e074dcd5bd213/SEED_AGE_5_REPORT_FEB.pdf (Accessed: 12 August 2025).

Tang, S., Bradley, E., Martin, K., Schwendel, G. and Styles, B. (2024) *Randomised controlled trial evaluation of the White Rose Maths Reception Jigsaw*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Reception-Jigsaw-addendum-report-final.pdf?v=1723591706 (Accessed: 12 August 2025).

The R Foundation (2025) 'The R project for statistical computing'. Available at: https://www.r-project.org/ (Accessed: 12 August 2025).

Zeileis, A. (2006) 'Object-oriented computation of sandwich estimators', *Journal of Statistical Software*, 16(9), pp. 1–16. Available at: https://doi.org/10.18637/jss.v016.i09.

Zeileis, A., Köll, S. and Graham, N. (2020) 'Various versatile variances: an object-oriented implementation of clustered covariances in R', *Journal of Statistical Software*, 95(1), pp. 1–36. Available at: https://doi.org/10.18637/jss.v095.i01.

Appendix A: R randomisation code

```
library(openxlsx)

#1. Set work directory
setwd("K:/EERJ/CfS/Randomisation")

#2.identify project
project<-"EERJ"

#3.identify classification: c, r or p
classification<-"c"

#4. Number of the randomisation: 1st, 2nd, 3rd ...
randomisation<-1
randomisation<-as.character(as.roman(randomisation))

#5. Load data
wb<-loadWorkbook("...")
sheets(wb)
Experiment<-read.xlsx(wb,sheet=1)

#Identify stratification and cluster variables

#6.list the stratification variables
stratification<-list("Geographic_region")
n_strats<-length(stratification)

#7.identify the cluster variable
cluster<-"NFER_No"

#8. What time is now? (hh.mm)
time_now<-16.32

aux<-100*trunc(time_now)+100*(time_now-trunc(time_now))
set.seed(aux)
seeds<-sample(1:9999,size=(n_strats+2))

#Keep the original order of the columns
originalColOrder<-colnames(Experiment)

###Adding a variable that will allow for the recovery
##of the original order of the data frame rows later on
Experiment$originalRowOrd<-1:nrow(Experiment)

### Ordering Experiment by cluster
Experiment<-Experiment[order(Experiment[,cluster]),]
```

```
### Assigning a random order to the stratification
rands<-paste("rand",as.character(1:n_strats),sep="_")

for (i in 1:n_strats){
 aux<-as.data.frame(sort(unique(Experiment[,stratification[[i]]])))
 set.seed(seeds[1])
 seeds<-seeds[-1]

 aux[rands[i]]<-sample(1:nrow(aux))

 Experiment<-merge(Experiment,aux,by.x=stratification[[i]],by.y=colnames(aux)[1])
}

###Randomise by cluster
set.seed(seeds[1])
seeds<-seeds[-1]
Experiment["rand_cluster"]<-sample(nrow(Experiment))

###Reorder the rows of Experiment by rands and rancluster
rands<-c(rands,"rand_cluster")
aux<-do.call(order,Experiment[rands])
Experiment<-Experiment[aux,]

###Assigning Control or Intervention Group
aux<-rep(1:2,times=round(nrow(Experiment)/2))
Experiment$grp<-aux[1:nrow(Experiment)]

aux<-data.frame(School_Randomisation_Group=c("control","intervention"))
set.seed(seeds[1])
aux$randgroup<-sample(1:2)

Experiment<-
merge(Experiment[,!colnames(Experiment)%in%"School_Randomisation_Group"],aux,by.x="gr
p",by.y="randgroup")

##Returning the data frame to its original order
Experiment<-Experiment[order(Experiment$originalRowOrd),]

###Removing the variables that are no longer necessary
rands<-c("originalRowOrd",rands)
rands<-which(colnames(Experiment)%in%rands)
Experiment<-Experiment[,-rands]
Experiment<-Experiment[,originalColOrder]
```