# Reciprocal Reading effectiveness trial
# Statistical Analysis Plan


## Evaluator (institution): Behavioural Insights Team
## Principal investigator(s): Patrick Taylor


Education Endowment Foundation

Template last updated: August 2019

| | |
|---|---|
| PROJECT TITLE[1] | Reciprocal Reading effectiveness trial |
| DEVELOPER (INSTITUTION) | FFT (Fischer Family Trust) |
| EVALUATOR (INSTITUTION) | Behavioural Insights Team |
| PRINCIPAL INVESTIGATOR(S) | Dr Patrick Taylor |
| PROTOCOL AUTHOR(S) | Neus Torres Blas, Dr Patrick Taylor |
| TRIAL DESIGN | Two-arm cluster randomised controlled trial with random allocation at school level |
| TRIAL TYPE | Effectiveness |
| PUPIL AGE RANGE AND KEY STAGE | Key Stage (KS) 2 Years 5 and 6 (Ages 9 to 11) |
| NUMBER OF SCHOOLS | 295 |
| NUMBER OF PUPILS | 4263 at randomisation |
| PRIMARY OUTCOME MEASURE AND SOURCE | New Group Reading Test (NGRT) overall test score |
| SECONDARY OUTCOME MEASURE AND SOURCE | NGRT passage comprehension score<br>NGRT sentence completion score<br>KS2 SATs reading score |

[1] Make sure that the project title here matches the title of the document and the protocol. Please ensure that there is an identification as a randomised trial in the title as per CONSORT requirements.

## SAP version history

| VERSION | DATE | REASON FOR REVISION |
|---|---|---|
| 1.2 [*latest*] | May 2025 | Deleted a typo in the missing data section and the CACE regression model. |
| 1.1 | March 2025 | Deleted a typo in the secondary outcome regression models. |
| 1.0 [*original*] | January 2024 | *N/A* |

# Table of Contents

## Introduction

Reciprocal Reading is a structured, discussion-based reading comprehension program developed by FFT. It targets Year 5 and 6 students who can decode text but struggle with comprehension. The program employs four key strategies: predicting, clarifying, questioning, and summarising. These are applied to small text sections to address comprehension issues in real-time. The intervention is delivered twice weekly for 12 weeks by trained teaching assistants or teachers in groups of 6–8 students.

The purpose of the upcoming analyses is to evaluate the effectiveness of Reciprocal Reading at scale, following a previous efficacy trial (O'Hare et al., 2019). This new evaluation aims to determine if the positive impacts observed in the earlier trial - such as the average of +2 months' additional progress in overall reading ability and reading comprehension and the larger gains for students eligible for free school meals - can be replicated in a larger, more diverse set of schools across England, under closer to real world conditions.

This effectiveness evaluation will be conducted as a two-armed cluster-randomised controlled trial involving approximately 300 schools. These schools have been randomly assigned to either implement the Reciprocal Reading program or continue with business-as-usual. The primary outcome of the evaluation will focus on the impact of the intervention on overall reading, measured using the NGRT. Secondary outcomes will assess reading accuracy and reading comprehension scores in the NGRT, and Year 6 KS2 reading scores. Additionally, the evaluation will examine the intervention's impact on pupils receiving free school meals and explore its potential as a tool for closing the attainment gap.

The Reciprocal Reading effectiveness trial was pre-registered in the OSF Framework in August 2023. The registration DOI is https://doi.org/10.17605/OSF.IO/8RHFD.

## Design overview

Table 1: Trial design overview

| | | |
|---|---|---|
| Trial design, including number of arms | | Two-arm, cluster randomised |
| Unit of randomisation | | School |
| Stratification variables (if applicable) | | Batched randomisation |
| Primary outcome | variable | **Reading proficiency (1)** |
| | measure (instrument, scale, source) | New Group Reading Test (NGRT), overall reading score, GL Assessment (1) |
| Secondary outcome(s) | variable(s) | Reading comprehension (2)<br>Sentence completion skill (3)<br>Reading attainment (4) |

| | | |
|---|---|---|
| Baseline for primary outcome | | |
| | measure(s) (instrument, scale, source) | NGRT passage comprehension score, GL Assessment (2)<br>NGRT sentence completion score, GL assessment (3)<br>KS2 SAT reading score (4) |
| | **variable** | **Reading proficiency** (1) |
| | measure (instrument, scale, source) | New Group Reading Test (NGRT) overall reading score, GL Assessment (1) |
| Baseline for secondary outcome | **variable** | Reading comprehension (2)<br>Sentence completion skill (3) |
| | measure (instrument, scale, source) | NGRT passage comprehension score, GL assessment (2)<br>NGRT sentence completion score, GL assessment (3) |

## Sample size calculations overview

Table 2: Sample size calculations at protocol and randomisation stage

| | | Protocol | | Randomisation | |
|---|---|---|---|---|---|
| | | OVERALL | FSM | OVERALL | FSM |
| **Minimum Detectable Effect Size (MDES) in Hedges' G** | | 0.13 | 0.16 | 0.13 | 0.16 |
| **Pre-test/ post-test correlations** | level 1 (pupil) | 0.56 | 0.58 | 0.56 | 0.58 |
| | level 2 (class) | - | - | - | - |
| | level 3 (school) | - | - | - | - |
| **Intracluster correlations (ICCs)** | level 2 (class) | 0.15 | 0.18 | 0.15 | 0.18 |
| | level 3 (school) | - | - | - | - |
| **Alpha**[2] | | 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 |

[2] Please adjust as necessary for trials with multiple primary outcomes, 3-arm trials etc. when a Bonferroni correction is used to account for family-wise errors.

| | | Protocol | | Randomisation | |
|---|---|---|---|---|---|
| | | OVERALL | FSM | OVERALL | FSM |
| **Average cluster size** | | 12 | 6 | 14 | 6 |
| **Number of schools** | intervention | 150 | 150 | 148 | 148 |
| | control | 150 | 150 | 147 | 147 |
| | **total** | 300 | 300 | 295 | 295 |
| **Number of pupils** | intervention | 1800 | 900 | 2118 | 829 |
| | control | 1800 | 900 | 2145 | 848 |
| | **total** | 3600 | 1800 | 4263 | 1677 |

The sample size calculations at the protocol stage were based on the following assumptions:

- A baseline correlation between baseline and endline primary outcome of 0.56 for the full sample ($R^2 = 0.56^2$) and 0.58 for the FSM subsample ($R^2 = 0.58^2$), as reported in the EEF efficacy trial (targeted intervention);
- Intracluster correlation of 0.15 for the full sample and 0.18 for the FSM subsample, as reported in the efficacy trial (targeted intervention);
- Pupil attrition of 15%, as reported in the efficacy trial (targeted intervention);
- School-level clustered randomisation with a cluster size equal to 12 pupils, with 2 delivery groups of 6 pupils per school, before attrition;
- A proportion of FSM pupils of 50%, for the subgroup analysis (a cluster size of 6 FSM-eligible pupils per school). The proportion of FSM pupils in the targeted intervention of the efficacy trial according to the EVERFSM_6_P variable in the NPD was 54%. We assumed a slightly smaller percentage of FSM-eligible pupils because we expect more variability in our sample. The national average rates of FSM pupils in Year 5 and 6 for 2022/23 were 26% and 25.4% according to the UK School Census;[3]
- Alpha of 0.05 and statistical power of 80%;
- Equal-sized treatment and control groups, where all pupils in the intervention and control groups will be tested;
- NGRT overall score mean of 288.46 and SD of 49.98 for the full sample, and a mean of 282.54 and SD of 49.70 for the FSM subsample, as reported in the efficacy trial (targeted intervention);
- A recruitment target of 300 schools.

[3] ONS. "Schools, pupils and their characteristics. Academic year 2022/23", 8 June 2023. Accessible at https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics

The resulting MDES with these assumptions was 0.13 in Hedges G, and 0.16 for the FSM subgroup analysis (see Table 2 above). For context, the effect sizes reported in the efficacy trial were 0.14 for the overall sample and 0.16 for the FSM subsample.

A first batch of 225 schools was randomised in July 2023 and a second batch of 70, in September of the same year. The total number of successfully recruited schools for the trial at that point was 295. 148 schools were allocated to the treatment group and 147 to control, with 2118 pupils in treatment and 2145 in control, making a total of 4263 pupils.[4]

The number of recruited schools was smaller than the target at protocol, but the average number of pupils per school was higher than anticipated. This explains the difference in sample sizes between protocol and randomisation.

After updating the power calculations with the actual number of schools and cluster sizes at randomisation, the MDES in Hedges' G is 0.13 for the primary analysis, and 0.16 for the subgroup analysis on FSM pupils, after attrition, which are the same as the MDESs reported in the Trial Protocol (to 2dp). The trial is powered to detect an effect size (ES) smaller than the ES reported in the efficacy trial for the overall sample (0.13<0.14), by a narrow margin. Given the lack of feasibility to increase the sample size, we hope to increase the statistical power of the primary outcome analysis by including baseline characteristics other than attainment that have been shown by the NGRT test provider to be correlated with the test scores (see section "Primary outcome analysis" for a more detailed description of the model). However, for the FSM subgroup analysis, it is only powered to detect an effect slightly bigger (0.164 > 0.160). Note that the proportion of FSM pupils at randomisation was 40%, lower than the assumption at protocol stage.

Given the trial design is clustered, the MDES was calculated using the effective sample size (ESS). If each cluster has size n (after attrition) and the ICC is $\rho$, the ESS of each cluster was calculated with the following formula:

$$ESS = \frac{n}{1 + (n-1)\rho}$$

The ESS for the treatment and control group was computed after summing up all cluster ESSs for each trial arm and used to obtain the MDES with the pwr.t2n.test function in R studio.

We also computed the MDES for the full sample and FSM subgroup analysis under different levels of pupil attrition (20% and 10%) to compare them with the main scenario, which assumes 15% of attrition following the efficacy trial. The MDES did not vary substantially under either of the two cases (see Table 3, scenarios A and C).

Finally, we also calculated the MDES for a scenario with pupil and school attrition (scenario D). 3 schools in the treatment group that were randomised in batch 1 decided to withdraw from the programme after the Summer, citing lack of staff resources to carry out the intervention. We excluded these 3 schools from the sample, and another 3 schools from the control group randomly selected, and recalculated the MDES to gauge the potential impact of the early withdrawals on the study's power. The resulting MDES for the main analysis did not vary, and the MDES for the FSM subgroup analysis increased by 1pp from scenario B.

[4] In October 2023, three schools that were randomly assigned to the treatment group in July 2023 notified FFT they were withdrawing from the programme, citing staff changes and insufficient resources, leaving a remaining sample of 292 schools.

Under all of these scenarios, the MDES for the primary analysis remains at 0.13 (2dp), which is less than the estimated effect in the efficacy trial. In the two worst case scenarios, the MDES for the subgroup analysis increases to 0.17 (2dp) which is slightly larger than the estimated effect in the efficacy trial. The trial therefore appears to be well-powered for the primary analysis, but possibly slightly unpowered for the subgroup analysis in the event of higher-than-expected attrition.

Table 3: MDES at different levels of attrition

| | | Pupil attrition | | | School and pupil attrition |
|---|---|---|---|---|---|
| | | A) 20% | B) 15% (main scenario) | C) 10% | D) 3 schools per arm + 15% pupil attrition |
| **MDES in Hedges' G, all pupils** | | 0.13 | 0.13 | 0.13 | 0.13 |
| **MDES in Hedges' G, FSM pupils** | | 0.17 | 0.16 | 0.16 | 0.17 |
| **Number of schools after attrition** | intervention | 148 | 148 | 148 | 145 |
| | control | 147 | 147 | 147 | 143 |
| | **total** | 295 | 295 | 295 | 289 |
| **Number of pupils after attrition** | intervention | 1694 | 1800 | 1906 | 1769 |
| | control | 1716 | 1823 | 1930 | 1790 |
| | **total** | 3410 | 3623 | 3836 | 3559 |

The power calculations for the protocol and randomisation stages were done using the pwr package in R studio.

## Analysis

All primary and secondary outcome analyses will be conducted on an intention-to-treat (ITT) basis to compare treatment and control arms and will include all pupils selected by participating schools to take part in the trial. This analysis will be done with complete cases only (observations missing one or more values will be dropped from the analysis). Analyses will be conducted in Stata or Rstudio, using 2-sided significance tests, at the 5% significance level.

The analysis requires to identify a group of pupils in control schools comparable to the pupils receiving the intervention in treatment schools. Schools were asked to nominate pupils to participate in the programme during the recruitment process (prior to being informed of their

trial arm allocation). This will allow us to identify the pupils who would have received the intervention, had their school been assigned to the treatment group.

However, asking schools to nominate pupils before knowing whether they will take part in Reciprocal Reading could lead them to deviate from business-as-usual. Normally, a school would do the selection after deciding to take part in the programme. The assumption that the treatment effect estimate will capture the difference between participating in the programme and doing business-as-usual could be violated if control schools start giving pupils additional reading support to make up for not receiving the intervention, or if treatment schools substitute usual reading curricular activities for Reciprocal Reading sessions. However, out-of-class support for struggling pupils in control schools may not constitute a violation of this assumption, so long as it is credible that this support would have occurred regardless of treatment assignment. We will examine whether participation in the Reciprocal Reading trial has impacted business-as-usual through two retrospective surveys, one for treatment and another for control schools, that will be part of the IPE. We will present a summary of the results of those surveys to contextualise the findings for the impact evaluation analysis.

In addition to the effect sizes, we will summarise the means and standard deviations of baseline and endline primary and secondary outcomes for each trial arm, together with histograms of baseline and endline data distributions.

We will use the variable "EVERFSM_6_P" in the NPD as the variable for FSM eligibility (this variable captures free school meals eligibility in the past 6 years).

The analysis will be carried out using either Stata or R studio, in the latest version available in the ONS SRS at the time of analysis. We will specify the software version we use in the final report.

### *Primary outcome analysis*

Analysis of the primary outcome, the NGRT overall score, will be carried out using an ordinary least squares (OLS) regression with clustered standard errors at the school level, to reflect the clustered design of the trial:

(Equation 1)

$$Y_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 Batch_s + \beta_4 X_{is} + \epsilon_{is}$$

Where:

- $Y_{is}$ is the endline NGRT overall test score for individual i, in school s;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreNGRT_{is}$ is the baseline attainment for individual i, in school s, measured through the baseline NGRT overall test score;
- $Batch_s$ is a binary indicator of the randomisation batch (1 or 2) of school s;
- $X_{is}$is a is a vector of pupil-level covariates including year group, gender, EAL status and, if applicable (see the note below) post-test assessment time for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

According to the NGRT Technical Guidance[5] age, gender and EAL status are consistently correlated with NGRT scores, so they will be included in the model as individual-level covariates. This will increase the precision of the treatment estimate by reducing the idiosyncratic variance, and make the model more robust in case of spurious and moderate imbalances[6] in these covariates after randomisation.To assess the robustness of this model, we will estimate two other models, one without these covariates and another with treatment status as the only covariate, and compare the results of the main specification (see the section "Additional analyses" below).

The time window to collect the endline NGRT will be kept as narrow as possible to minimise the chance of differences in assessment timings between treatment and control schools, given that an additional month of schooling can already make a difference in academic attainment for children at this age. For reference, endline assessments in the efficacy trial were done between May and July. However, since the effectiveness trial is being carried out in 295 schools, the endline assessments will start earlier than this due to capacity constraints from the organisation carrying out the assessments, so the testing window will fall between the 18th of March and the end of the school year in July. Starting earlier could mean that pupils have less Reciprocal Reading sessions before the endline assessment than the participants in the efficacy trial, and so the average treatment effect might be smaller. To minimise this risk, all assessments will be conducted at least three weeks after a school has had the 2nd training session.

If there are more than 2 months of difference between the first and the last endline assessment, we will include assessment month fixed effects as an additional covariate in the regression to capture the variance in outcomes due to the date when the assessment was conducted. In this case, we will also include the assessment month in the covariates that are assessed for imbalance at baseline.

### *Secondary outcome analysis*

#### *a. NGRT passage comprehension subscale*

Analysis will be carried out using an ordinary least squares (OLS) regression with clustered standard errors at the school level, to reflect the clustered design of the trial:

(Equation 2)

$$PC_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PrePC_{is} + \beta_3 Batch_s + \beta_4 X_{is} + \epsilon_{is}$$

Where:

- $PC_{is}$ is the endline NGRT passage comprehension score for individual i, in school s;

[5] According to the test developers (GL Assessment), female pupils perform better in the NGRT test than male pupils by an average of 3.1 SAS (Standard Age Score) points, and non-EAL pupils perform better than EAL students by an average of 3.6 SAS points. These differences are significant at all age groups. See: https://support.gl-assessment.co.uk/knowledge-base/assessments/ngrt-support/general-information/technical-guidance/#:~:text=Gender%20differences,in%20primary%20and%20secondary%20schools (last accessed: 4th of February 2024).

[6] We will consider there is a moderate imbalance in a covariate when the normalised difference in means between treatment and control is between 5 and 10%, according to the EEF Evaluation Guidelines (2019). See the section "Imbalance at baseline" in this SAP.

- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- *Pre*$PC_{is}$ is the baseline NGRT passage comprehension score for individual i, in school s;
- *Batch* $_s$ is a binary indicator of the randomisation batch (1 or 2) of school s;
- $X_{is}$is a is a vector of pupil covariates including year group, gender, EAL status and, if applicable, post-test assessment time for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

The estimation of $\beta_1$ in this model may be biased because of sample selection. Pupils who do not score the minimum required points in the SC (sentence completion) section of the test will not be given PC (passage comprehension) tasks, so missing data on NGRT PC will not be random. Since the treatment is expected to improve SC abilities, students in the treatment group are more likely to achieve this improvement at the endline, and hence more likely to have a PC score. A more detailed explanation of this issue can be found in the Evaluation Protocol (pages 18-19).

The severity of the selection problem will be assessed with the following additional analyses:

a) reporting the percentage of pupils with an NGRT overall reading score that are missing the PC score, at baseline and endline[7];
b) reporting the number of pupils in the treatment and control groups that have no PC score at baseline but do at endline;
c) a logit model to test whether the probability of not having an endline score is correlated with treatment assignment (see equation 3 below).

(Equation 3)

Missingness of NGRT passage comprehension score will be modelled as follows:

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 T_{is} + \beta_2 preSC_{is}$$

where:

- $M_{is}$ is the binary variable for missingness (equal to 1 if NGRT PC is missing at endline and 0 if not missing);
- $p_{is}$ is the probability that a given observation is missing the NGRT PC score at endline;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $preSC_{is}$ is the baseline NGRT sentence completion score for individual i, in school s.

If more than 5% of pupils with NGRT overall score at endline are missing a PC score at endline, and the coefficient of treatment assignment in the logit model is negative and statistically significant at 5%, we will carry out further analysis. In this case, we will compare

[7] In the efficacy trial, 1% of the sample of pupils for the targeted intervention had overall reading score but not reading comprehension score at endline (O'Hare et al., 2019, Appendix H).

the results of estimating $\beta_1$ in Equation 2 for the full sample and with the subsample of pupils who were above the threshold to have a passage comprehension score before and after the trial.

### b. NGRT sentence completion subscale

Analysis of the NGRT sentence completion score will be done using the same model for the other two NGRT outcomes:

(Equation 4)

$$SC_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreSC_{is} + \beta_3 Batch_s + \beta_4 X_{is} + \epsilon_{is}$$

Where:

- $SC_{is}$ is the endline NGRT sentence completion score for individual i, in school s;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreSC_{is}$ is the baseline NGRT sentence completion score for individual i, in school s;
- $Batch_s$ is a binary indicator of the randomisation batch (1 or 2) of school s;
- $X_{is}$is a is a vector of pupil covariates including year group, gender, EAL status and, if applicable post-test assessment time for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

### c. KS2 English reading SATs for Year 6 pupils

We will also estimate the impact of the intervention on the KS2 SAT reading scores of the subsample of pupils that are in Year 6 at the start of the programme and do their KS2 SATs in May 2024.

The estimation model is equivalent to the model used for the primary outcome:

(Equation 5)

$$KS2READSCORE_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 Batch_s + \epsilon_{is}$$

Where:

- $KS2READSCORE_{is}$ is KS2 reading attainment scaled score, for individual i, in school s;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreNGRT_{is}$ is the baseline NGRT overall reading score for individual i, in school s;
- $Batch_s$ is a binary indicator of the randomisation batch (1 or 2) of school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

The variable KS2_KS2READSCORE will be obtained from the KS2 dataset from the NPD. This is the scaled score on the English reading test, and can range between 59 to 120, where 100 indicates the pupil met the expected standard of the test.[8]

We will be limited in our ability to draw causal conclusions from this subgroup analysis, as it is likely to be underpowered. The sample for this model will have the same number of clusters as the main analysis, but cluster size will be around 50% smaller, which is the same scenario assumed in the power calculations for the FSM subgroup analysis. This means we will only be powered to detect a causal effect in Year 6 pupils equal or higher to 0.16 SD, while the effect size in the targeted intervention of the efficacy trial was 0.14 SD. However, it will provide indicative evidence of the impact of Reciprocal Reading on standardised reading attainment measures.

### *Subgroup analysis for FSM-eligible pupils*

We will do two subgroup analyses for FSM-eligible pupils:

1. Estimate the model specified in the primary analysis on the subsample of FSM-eligible pupils; (see Equation 6 below).
2. Estimate a similar model including an interaction term for FSM eligibility, using a pooled sample (see Equation 7 below).

Both approaches will estimate the effect size for FSM pupils, but the latter uses information from the whole sample. Under ideal conditions, the total treatment effect in both should be analogous. The results for both will be compared and reported.

FSM eligibility will be derived from the variable "EVERFSM_6_P" in the NPD.

#### *Model 1: Split sample estimation*

(Equation 6)

$$[Y_{is} = (\delta_0 + \delta_1 T_{is} + \delta_2 PreNGRT_{is} + \delta_3 Batch_s + \delta_4 X_{is} + \epsilon_{is})] \mid EverFSM_{is} = 1$$

Where:

- $Y_{is}$ is the endline NGRT overall test score for individual i, in school s;
- $T_{is}$ is a binary indicator for the treatment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreNGRT_{is}$ is the baseline attainment for individual i, in school s, measured through the baseline NGRT overall test score;
- $Batch_s$ is a binary indicator of the randomisation batch (1 or 2) of school s;
- $X_{is}$is a is a vector of pupil covariates including FSM status, year group, gender, EAL status and, if applicable, post-test assessment time for individual i, in school s; and

[8] We are opting for KS2_KS2READSCORE instead of KS2_READSCORE because we expect the number of missing values to be smaller. In the case of KS2_READSCORE, which ranges from 80 to 120, it could be missing for 3-4% of pupils in the sample due to not meeting the minimum test standard, according to the national average. KS2_KS2READSCORE has a wider range and less missing observations as notional values are assigned to pupils who are not tested based on teacher assessment.

- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

This model will be estimated for the subsample of pupils with *EverFSM*$_{is}$ equal to 1.

The reported effect size for FSM-eligible pupils will be the treatment coefficient in Equation 6 ($\delta_1$), in Hedges G.

#### *Model 2: Interaction effect*

(Equation 7)

$$Y_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 EverFSM_{is} + \beta_3 (EverFSM_{is} * T_{is}) + \beta_4 PreNGRT_{is} + \beta_5 Batch_s + \beta_6 X_{is} + \epsilon_{is}$$

Where:

- *Y*$_{is}$ is the endline NGRT overall test score for individual i, in school s;
- *T*$_{is}$ is a binary indicator for the treatment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- *EverFSM*$_{is}$ is a binary indicator equal to 1 if the pupil has been eligible for FSM in the past 6 years and 0 if not;
- *PreNGRT*$_{is}$ is the baseline attainment for individual i, in school s, measured through the baseline NGRT overall test score;
- *Batch* $_s$ is a binary indicator of the randomisation batch (1 or 2) of school s;
- $X_{is}$is a is a vector of pupil covariates including FSM status, year group, gender, EAL status and, if applicable, post-test assessment time for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

The interaction term coefficient $\beta_3$, its SD and p-value will be reported, emphasising the range of effects that are compatible with the 95% confidence interval.

As a sensitivity check, we will also compute the ES for FSM-eligible pupils from Model 2 and compare it with Model 1. The ES from Model 2 will be calculated as the sum of $\beta_1$ (the coefficient from the ITT variable) and the interaction effect $\beta_3$. This sum should be analogous to the treatment coefficient found in the split sample Model 1, $\delta_1$.

$$ES\ for\ FSM = \beta_1 + \beta_3 = \delta_1$$

We will explore any difference and discuss its implications for the results in the Annex.

### *Additional analyses*

#### *Models without covariates*

All primary outcome and secondary outcome models will be re-estimated with the treatment assignment as the only covariate. The results of the analyses without covariates will be compared to the results from the main specification.

Additionally, the primary outcome will also be re-estimated with a model without pupil-level covariates, only the treatment assignment, baseline attainment and trial design characteristics

(the randomisation batch). This model will allow for more comparability with other EEF trials and will be used to see if the inclusion of individual-level covariates in the main specification changes the results meaningfully.

(Equation 8)

$$Y_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 Batch_s + \epsilon_{is}$$

Where:

- $Y$is is the endline NGRT overall score for individual i, in school s;
- Tis is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- Pre$NGRT$is is the baseline NGRT overall score for individual i, in school s;
- Batch s is a binary indicator of the randomisation batch (1 or 2) of school s; and
- εis is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

### *Longitudinal follow-up analyses*

None.

### *Imbalance at baseline*

Imbalance at baseline will be assessed with a cross-tabulation of school and pupil-level background characteristics. See Table 8 below (extracted from the EEF Evaluation report template) for the template of the output that we will produce.

We will assess balance at baseline by comparing normalised differences[9] in means of baseline test scores and demographic covariates correlated with the outcomes (year group, gender, FSM status, EAL status and, if applicable, post-test assessment time) for all units as randomised (before attrition) and as analysed (after attrition). We will also assess balance at the points of baseline and analysis on key school characteristics (urban/rural, Ofsted rating, and proportion of FSM pupils, KS2 SAT reading performance) The normalised differences will be computed as effect sizes in Hedges G, following the guidelines in the "Effect size calculation" section below (they will be equal to the difference in control and treatment means divided by the pooled SD).

If an imbalance higher than 0.05 SD is detected, we will assess whether this is likely to be due to chance or if there were any issues with the randomisation process. We will also run sensitivity analyses by including the variables where imbalance was found and seeing if the results for the primary outcome change. We will then discuss the implications for the security of the findings.

Table 4: Baseline characteristics of groups as randomised *(analysed in a separate table)*

[9] The normalised difference is defined as the difference in means between the two groups, divided by the pooled standard deviation. Normalised differences with a magnitude of 0.1 or less indicate a negligible correlation between the covariate and assignment to treatment group, which can usually be addressed through covariate adjustment in the regression (Austin, 2009). We will use the EEF's stricter threshold for risk and consider as a potentially meaningful risk any imbalance higher than 0.05 SD, according to EEF (2019), "Classifying the security of EEF findings".

| School-level (categorical) | National-level mean | Intervention group | | Control group | | |
|---|---|---|---|---|---|---|
| | | n/N (missing) | Count (%) | n/N (missing) | Count (%) | Normalised difference (pooled SD) |
| Location<br>- Urban<br>- Rural | | | | | | |
| Ofsted rating<br>- Outstanding<br>- Good<br>- Requires Improvement<br>- Serious weakness | | | | | | |
| School-level (continuous) | | n/N (missing) | Mean (SD) | n/N (missing) | Mean (SD) | Normalised difference (pooled SD) |
| Proportion of FSM pupils | | | | | | |
| KS2 SAT reading performance | | | | | | |
| Pupil-level (categorical) | | n/N (missing) | Count (%) | n/N (missing) | Count (%) | Normalised difference (pooled SD) |
| Year group<br>- Year 5<br>- Year 6 | | | | | | |
| Gender<br>Male<br>Female<br>Other | | | | | | |
| FSM | | | | | | |
| EAL | | | | | | |
| Pupil-level (continuous) | | n/N (missing) | Mean (SD) | n/N (missing) | Mean (SD) | Normalised difference (pooled SD) |
| Baseline NGRT overall score | | | | | | |
| Baseline NGRT PC score | | | | | | |
| Baseline NGRT SC score | | | | | | |
| Post-test assessment month | | | | | | |

### *Missing data*

#### *Descriptive analysis*

For FSM eligibility, gender, EAL status and KS2 SAT attainment, we will use data from the NPD (as opposed to data from schools) where possible to minimise missingness.

We will do cross-tabulations to report the number of missing observations for the following cases:

A. Missing covariates (gender, EAL status, baseline NGRT scores), for the primary and secondary outcome analyses

B. Number of complete cases, for all outcome analyses

C. Missing outcome data, for the sample at randomisation, for the treatment and control groups respectively

See table below for an example.

Table 5: Missing covariates

| Outcome | Covariate 1 | | | Covariate 2 | | | Complete cases n/N (%) |
|---|---|---|---|---|---|---|---|
| | Treatment n/N (%) | Control n/N (%) | Total n/N (%) | Treatment n/N (%) | Control n/N (%) | Total n/N (%) | |
| NGRT overall score | | | | | | | |
| NGRT passage comprehension score | | | | | | | |
| NGRT sentence completion score | | | | | | | |
| KS2 Reading score | | | | | | | |

For all outcomes, we will compare the treatment estimates of the complete case analysis with the treatment estimate of an unadjusted model which will have treatment assignment as the only covariate. The results from the primary outcome model will also be compared to the results from an alternative specification that will not include year group, gender and EAL status. These comparisons are specified in more detail in the "Additional analyses" section, "Models without covariates". No further analysis will be conducted for the secondary outcomes.

#### *Understanding patterns of missingness*

For the primary outcome analysis, where covariates are missing for more than 5% of the sample with primary outcome data, or the primary outcome is missing for more than 5% of the randomisation sample, we will first try to identify the pattern of missingness, i.e. whether they are missing conditional on other covariates or outcomes, or not.

Data is missing completely at random (MCAR) when missingness is uncorrelated with both observables and unobservables. This could occur in the case of pupils missing a test because they were sick, or because they left the school. Whether missing data is correlated (or not) with unobservables will depend on the context of the trial. Whenever possible, we will try to gather information from the schools on the reason for a missing test result during baseline and endline data collection and try to identify whether it was a case of persistent or a one-time absence, a withdrawal from the trial or the evaluation, or that the pupil left the school. In the case of MCAR, complete case analysis will give unbiased results, but will have less statistical power.

Data is missing at random (MAR) when missingness is correlated with other covariates, and missing not at random (MNAR) when missingness is correlated with unobservables. In both cases, the complete case analysis will give biased results. The analysis approach will depend on the type of missingness and whether the missing data are covariates or outcome variables.

In the case of covariates, we do not expect high levels of missing data for covariates obtained from the NPD. As for the baseline attainment, it is likely that some observations will be missing values due, for example, to pupils in the sample not sitting the tests, persistent absence, or pupils leaving the school before they had a chance to sit the test. In the trial, 88% of schools (all 225 schools from batch 1 and 35 out of 70 schools from batch 2) sat the baseline tests before knowing their treatment allocation, so data missingness on baseline attainment is unlikely to cause bias in the treatment estimate from the complete case analysis.

The missing data patterns in the primary outcome will be identified using the following logistic regression models to predict missingness:

(Equation 8)

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 T_{is} + \beta_2 preNGRT_{is} + \beta_3 X_{is}$$

where:

- $M_{is}$ is the binary variable for missingness (equal to 1 if the outcome is missing and 0 if not missing);
- $p_{is}$ is the probability that a given observation is missing the KS2 maths score;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $preNGRT_{is}$ is baseline reading attainment for individual i, in school s (the measure of baseline reading attainment will vary to match the missing outcome that is being modelled); and
- $X_{is}$ is a vector of pupil-level covariates including FSM status, gender and EAL status.

Missingness of covariates will be modelled as follows:

(Equation 9)

$$M_{is} \sim binomial(p_{is});\ logit(p_{is}) = \beta_0 + \beta_1 T_{is} + \beta_2 Y_{is} + \beta_3 X_{is}$$

where:

- $M_{is}$ is the binary variable for missingness (equal to 1 if the baseline reading attainment measure is missing and 0 if not missing);
- $p_{is}$ is the probability that a given observation is missing the KS2 maths score;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $Y_{is}$ is outcome for individual i, in school s (the endline outcome measure will vary to match the missing baseline measure that is being modelled); and
- $X_{is}$ is a vector of pupil-level covariates including FSM status, gender and EAL status.

We may also include available school-level covariates or interactions between covariates and treatment status if there are any indications during the trial that they may be correlated with missingness. For example, the schools that were randomised in the second batch had less time to conduct the baseline NGRT assessments and any mop-ups for pupils that were not at school on the day of the test, so they might present higher levels of missing data.

P-values below 0.05 will be considered evidence of missingness being conditional on covariates or MAR (missing at random). In this case, complete case analysis may yield biased estimates that will be corrected using multiple imputation, using the findings from this analysis. If there is no evidence that data is missing conditional on observables, data missingness may be MCAR (missing completely at random) or MNAR (missing not at random). The approach we take to account for the missing data will vary depending on the type of data missingness we find. The following section outlines the robustness checks we will perform in each of these three cases.

### *Robustness checks in the case of missing covariates*

If covariates are missing for more than 5% of the sample with primary outcome data, in addition to comparing the complete case analysis with the unadjusted model, we will run the following robustness checks:

1. If covariates are MCAR, a missing-indicator method (also known as MIM) will be used to increase the precision of the treatment estimate and compare to the complete case analysis. The MIM consists of creating an indicator of missingness, and using this to re-estimate the effects (see equation 10 below).
2. If covariates are missing conditional on other covariates (MAR),[10] we will re-run the primary outcome analysis using Multiple Imputation (MI). The approach to multiple imputation is specified below.
3. If covariates are MNAR (missingness is correlated with unobservables), multiple imputation will not correct for the bias. In this case, we may choose to use sensitivity analysis that will consist of re-estimating the model after excluding the relevant covariate and comparing the results with the main estimates. This will provide an indication of the direction of the potential bias.

**Missingness-indicator method**

The primary outcome analysis with the missingness-indicator would be modified as follows:

(Equation 10)

[10] Schultz and Grimes (2002) suggest that, when less than 5% of data is missing, there is likely to be little bias introduced to estimated treatment effects, so we have adopted this threshold here.

$$Y_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 X_{is} + \beta_4 M_{is} + \epsilon_{is}$$

where:

- $Y_{is}$ is the endline NGRT overall test score for individual i, in school s;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreNGRT_{is}$ is the baseline attainment for individual i, in school s, measured through the baseline NGRT overall test score;
- $X_{is}$is a is a vector of pupil and school-level covariates including FSM status, year group, gender, EAL status, and randomisation batch for individual i, in school s;
- $M_{is}$is a is a vector of missing indicators for individual-level covariates, equal to 1 if the covariate is missing and 0 otherwise, for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level.

In the MIM, all missing covariate observations are recoded to an arbitrary constant so the observations are not dropped by Stata/R.

If the covariate missingness is completely at random and unrelated to the outcomes or other covariates, this method will yield unbiased estimates and will be more efficient than the complete-case analysis (Zhao and Ding, 2022).

**Multiple imputation**

If a covariate is MAR conditional on other covariates and is missing for more than 5% of the sample used in the primary outcome model, we will use multiple imputation in line with EEF guidelines.

We will use predictive mean matching to impute missing values of the baseline NGRT within each dataset and generate at least 20 datasets. The estimated coefficients and standard errors will be pooled across the imputed datasets using Rubin's rules (Rubin, 1987). The results of the MI model will be compared to the main results of the primary analysis with complete cases.

***Robustness checks for missing outcome data and attrition***

If more than 5% of the sample at randomisation is missing primary outcome data, the additional analysis will depend on the type of missingness.

If outcome data is missing MCAR, the results of the main specification (which is estimated with complete cases) will give unbiased estimates, but the estimation will be less precise. The implications for the robustness of the results will be discussed in the report.

If the outcome is MAR conditional on identified and available covariates that are not in the main specification, we will estimate a model including those covariates and interpret the results. We will also estimate an unadjusted model and compare the results with the other models.

If the primary outcome data is MNAR and there is indicative evidence of differential attrition,[11] we will use bounds analysis (Manski-type bounds or Lee bounds) to determine an

[11] The three schools that withdrew from the intervention between the randomisation and the writing of this SAP were in the treatment arm and cited lack of resources after staff changes and recruitment

interval for the true treatment effect that corrects for bias from differential attrition. The use of Lee bounds (Lee, 2009) is preferred over extreme value or "Manski" bounds (Horowitz and Manski, 1998), but the use of Lee bounds will depend on whether we are confident that the monotonicity assumption holds (Lee, 2009).

### *Compliance*

Compliance for this analysis will be defined at the pupil level. We will consider a pupil in the treatment arm as compliant if they attended at least 20 sessions, based on FFT's recommendation.[12] Any amount below that will be considered as non-compliance.

Compliance will be defined by a binary variable equal to 1 if a pupil attended at least 20 Reciprocal Reading sessions, and 0 otherwise. This will be based on pupil attendance data that will be collected routinely by the teachers leading the sessions. The completion of pupil registers by teachers will be routinely monitored by FFT, who will share the data with BIT at the end of the intervention.

We will estimate the Complier Average Causal Effect (CACE) on the primary outcome using the following two stage least squares (2SLS) estimation model:

Stage 1 (Equation 11)

$$Z_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 X_{is} + \epsilon_{is}$$

Stage 2 (Equation 12)

$$Y_{is} = \beta_0 + \beta_1 \hat{Z}_{is} + \beta_2 PreNGRT_{is} + \beta_3 X_{is} + \epsilon_{is}$$

Where:

- $Z_{is}$ is the binary compliance indicator for individual i, in school s;
- $T_{is}$ is a binary indicator of the treatment assignment for individual i, in school s (1 if the pupil is in a treated school and 0 if not);
- $PreNGRT_{is}$ is the baseline attainment for individual i, in school s, measured through the baseline NGRT overall test score;
- $Y_{is}$ is the endline NGRT overall test score for individual i, in school s;
- $\hat{Z}_{is}$ are the predicted levels of compliance from the 1st stage of the 2SLS of Equation 11;
- $X_{is}$ is a vector of pupil and school-level covariates including year group, gender, EAL status, randomisation batch and, if applicable, post-test assessment time for individual i, in school s; and
- $\varepsilon_{is}$ is the cluster-robust error term, for individual i in school s, clustered at the school level (assuming the errors are correlated within school and reflecting the design of the study).

issues during the Summer. Reciprocal Reading is an intensive programme that requires schools to have two weekly sessions led by teachers and TAs, so we are expecting this type of attrition to affect disproportionately more treatment schools during the intervention. It is also possible that some control schools also withdraw from the evaluation activities because of resource constraints, but the demands of the evaluation for control schools are substantially smaller than delivering the programme.

[12] For reference, the minimum length of the intervention recommended to schools by FFT is 12 weeks of delivery, and two sessions per week. This makes 24 sessions in total. A minimum of 20 includes a margin for pupils to miss up to 4 sessions.

The results from the 1st and 2nd stage will be reported, together with the correlation between the instrument (the treatment assignment) and the endogenous variable (compliance indicator).

This model assumes that treatment does not have an effect for non-compliers. However, there may be an effect for pupils that have completed a number of sessions just below the 20-session limit. In order to test the robustness of the IV estimate, we will re-estimate the model using different thresholds that fall just below 20 (18, 15) and check if results change significantly.

We will do a second sensitivity analysis where we will consider a pupil as compliant if they did the 20 sessions in a maximum of 12 weeks (excluding school holidays). This will exclude from the complier group those schools and pupils that spaced out the sessions beyond the frequency recommended by FFT, as reducing the frequency of practice could dilute the treatment effect. The observance of this has been speculated upon in similar EEF trials (in prep).

### *Optimal dosage analysis*

We also wish to investigate the optimal number of instructional sessions required to maximise gains in reading comprehension, up to the point where additional sessions yield progressively smaller improvements. This analysis will aim to explore if and when the benefit from extra sessions decreases.

Ideally, the investigation into the optimal dosage would involve the random assignment of session frequencies across treatment schools to support an unbiased estimate of the impact of varying instructional intensities. As session frequency has not been randomly assigned, a descriptive analysis will be employed instead. The analysis will systematically examine the relationship between the number of sessions and their impact on reading comprehension outcomes. It seeks to identify a potential threshold beyond which the incremental benefits of additional sessions diminish. This will enable educators to empirically determine the most effective number of sessions to implement, based on a detailed understanding of how session frequency correlates with reading comprehension improvements.

The following two limitations will prevent us from identifying a causal relationship between the number of sessions and the change on overall reading ability:

1. There could be a low variability in the number of NGRT sessions across schools due to the scheduling constraints of the post-test assessments. Because of the high number of schools to be tested before the end of the school year, treatment schools will have to be tested as soon as they have done their 2nd training session and an additional three weeks of sessions to assimilate the training content. This means there will be many schools that will be tested when they have held a similar number of sessions, which will reduce the range of treatment levels (i.e., the number of Reciprocal Reading sessions) across the study's participants. This will limit our ability to see how the NGRT score varies with different levels of sessions.
2. Students with higher ability could self-select into doing more NGRT sessions (or less able students could prefer to miss out on more sessions). In this case, the number of sessions would be endogenous to the outcome, and the relationship between them could not be interpreted as causal.

To assess the degree of endogeneity, we will begin by regressing the number of sessions on pupil individual characteristics (including baseline attainment). If a regression coefficient is

statistically significant at the 5% level, that individual characteristic will be considered a predictor of session attendance and evidence of endogeneity. These findings will be used to moderate the causal interpretation of this analysis.

Then we will regress the change between baseline and endline NGRT score on a set of binary indicators for the number of sessions taken and on individual characteristics (gender, FSM status, EAL and baseline NRGT score). The binary indicators w1, w2, …wn will be equal to 1 if a pupil took a number of sessions that was in the nth percentile, and 0 otherwise, with the reference category being no sessions attended. The percentiles will be decided after obtaining the data on the distribution of sessions in order to have a similar number of pupils in each category while remaining informative.

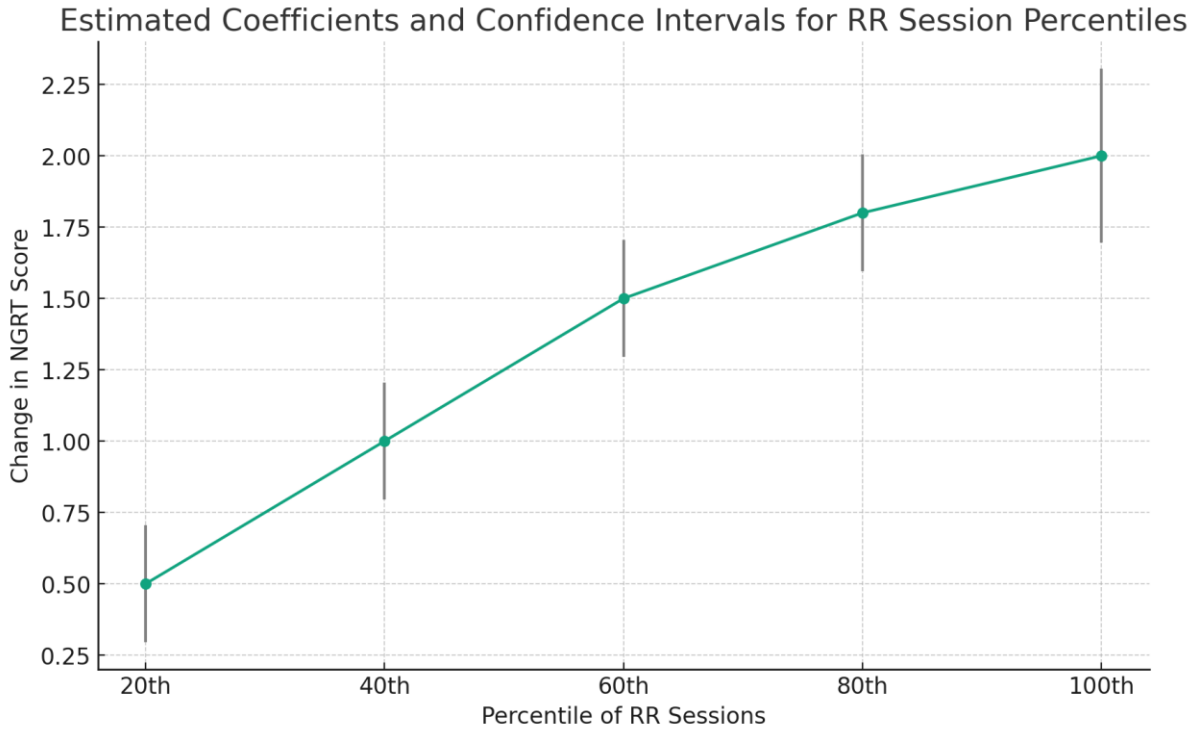
Equation 13:

$$\Delta NGRT_i = \beta_0 + \beta_1 W_i + \beta_2 PreNGRT_i + \beta_3 X_i + \epsilon_i$$

Where:

- $W_i$ is a vector of binary indicators $w_1, w_2, \ldots w_n$ of the number of sessions for individual i;
- $PreNGRT_i$ is the baseline attainment for individual i, measured through the baseline NGRT overall test score;
- $\Delta NGRT_i$ is the change between baseline and endline NGRT overall test score for individual i;
- $X_{is}$is a vector of pupil and school-level covariates including FSM status, gender and EAL status; and
- $\varepsilon_i$ is the heteroskedasticity-robust error term, for individual i.

The estimated coefficients for the W binary indicators will suggest how much change in NGRT score is associated with being in that percentile of session attendance compared to attending no sessions, controlling for individual characteristics. The last step will be to graph the change in NGRT on the coefficient estimates and their confidence intervals. See the graph below for an example:




Each dot represents the estimated coefficient for the change in NGRT score associated with taking the number of lessons in that percentile, with lines connecting them to show the trend. The vertical lines represent the confidence intervals for each coefficient estimate, providing a sense of the statistical uncertainty around these estimates. This visualisation will help illustrate how changes in NGRT scores might be associated with different levels of session attendance, controlling for other individual characteristics. For example, if the coefficients of the 80th and 90th percentiles are similarly sized and their confidence intervals overlap, this could be an indication that taking more lessons beyond the 80th percentile did not lead to a meaningful difference for the NGRT scores. In that case we can use a t-test to see if two coefficients are statistically different or not, with the caveat that the small sample size in each percentile will probably make the confidence intervals too wide for a valid inference.

The percentile with the highest coefficient estimate will be considered as the optimal dosage, unless there is indicative evidence that pupils from a lower percentage achieved a similar change in NGRT scores. This optimal dosage analysis will present descriptive evidence that will be used to contextualise the findings for the CACE analysis and provide recommendations for FFT and schools on the recommended intervention dosage.

### *Intra-cluster correlations (ICCs)*

ICCs for the primary outcome (NGRT overall score) will be calculated at the school level at pre-test and post-test. We will also calculate the ICC for the KS2 SAT reading scores at post-test, as KS2 SAT scores will only be available for Year 6 pupils.

We will estimate a one-way random effects ANOVA model with the school as a random effect (see Equation 13 below). The ICC will be calculated from the different variance components derived from the model

(Equation 13)

$$Y_{is} = \mu + a_s + \varepsilon_{is}$$

Where:

- $Y_{is}$ is the outcome for individual i, in school s;
- $\mu$ Is the unobserved population mean;
- $a_s$ is the school random effect for school s; and
- $e_{is}$ is the random error effect for individual i in school s.

In this model the variance of $a_s$ is denoted $s^2_\alpha$ (the outcome variance at the school level) and the variance of $e_{is}$ is denoted $s^2_e$ (the residual variance). The school ICC will be defined as:

$$ICC = \frac{s_a^2}{s_a^2 + s_e^2}$$

which can be interpreted as the proportion of the total variance that is attributable to differences between schools. The value of the ICC indicates the extent to which the observed variability in test scores is due to variability between schools. A higher ICC suggests that the school a pupil attends has a greater influence over the outcome.

### *Effect size calculation*

The effect size for the primary and secondary analysis, including the FSM subgroup analyses, will be presented in terms of Hedges G using the following formula:

$$Hedges\ G\ = \frac{(\underline{Y}_T - \underline{Y}_c)_{adjusted}}{sd*}$$

where $(\underline{Y}_T - \underline{Y}_c)_{adjusted}$is the difference in conditional means of treatment and control, obtained from the estimation of the analysis model with covariates, and sd* (the pooled SD) is a weighted average of the unconditional SD of the outcome in treatment and control. sd* is calculated with the following formula:

$$sd^* = \sqrt{\frac{(n_T - 1)sd_T^2 + (n_C - 1)sd_C^2}{n_T + n_C - 2}}$$

where:

- $n_T$is the number of individuals in the treatment group that are included in the relevant outcome analysis, and $n_C$ is the same for the control group;
- $sd_T$ is the unconditional SD of the outcome for the subsample of individuals in the treatment group included in the relevant outcome analysis, and $sd_C$ is the same for the control group.

The effect size for the primary analysis and FSM-subgroup analysis will also be converted into months of progress for the evaluation report using the conversion table in the EEF report template.

We will report 95% confidence intervals for the effect sizes of all primary and secondary outcome analyses. The coefficients will also be reported together with one to three stars to show confidence of the estimate at different standard levels of significance (1%, 5%, 10%).

The passage comprehension and sentence completion subscales will be obtained from the same test as the primary outcome and measure two important components of reading attainment, which increases the likelihood of family-wise error rate for the two comparisons. Consequently, the p-values reported for the two secondary analyses on NGRT Sentence Completion and NGRT Passage Comprehension scales will be adjusted for multiple comparisons using the Benjamini-Hochberg method.

By contrast, we will not adjust the p-value of the other secondary outcome (KS2 reading results). The analysis on KS2 reading test results will be done with a sample subgroup so it is not an equivalent comparison. The interpretation of the results for this outcome will be distinct from the two NGRT subscales.

## References


Austin, P.C. (2009) Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research, Communications in Statistics - Simulation and Computation, 38:6, 1228-1234.

Education Endowment Foundation (2019). Classifying the security of EEF findings.

Education Endowment Foundation (2022). Statistical analysis guidance for EEF evaluations.

Horowitz, J. & Charles F. Manski (2000). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data, Journal of the American Statistical Association, 95:449, 77-84.

Lee, David. S. (2009) Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects, *The Review of Economic Studies*, Volume 76, Issue 3, 1071–1102.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. The American Economic Review, 80(2), 319–323.

O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., Biggart, A., Connolly, P., & Bower, C. (2019). Reciprocal Reading Evaluation Report. EEF project reports.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York.

Schaffer, JL. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research. 9(1):3-15. doi:10.1177/096228029900800102

Zhao, A., & Ding, P. (2022). To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. Journal of the American Statistical Association, 1-11.