Education Endowment Foundation

Reciprocal Reading effectiveness trial

Evaluation report

February 2026

Neus Torres Blas, Emma Forsyth, Andrés Cueto, and Patrick Taylor

BIT

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, colleges, and early years settings to improve teaching and learning for 2–9-year-olds through better use of evidence.

We do this by:

- **Summarising evidence**. Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds. Putting evidence to use.
- **Putting evidence to use**. Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were reendowed with an additional £137m from government, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:

The Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP

0207 802 1653

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Table of contents

About the evaluator

This project was evaluated by a team at the Behavioural Insights Team. The project was led by Neus Torres Blas with oversight from Dr Patrick Taylor. Neus Torres Blas conducted all quantitative research. The implementation and process evaluation was led by Emma Forsyth with support from Andrés Cueto.

Contact details:

The Behavioural Insights Team
58 Victoria Embankment,
London,
EC4Y 0DS

Email: info@bi.team

Acknowledgements

Fischer Family Trust Education: Katie Kielty (Education Product Manager); Laura James (Customer Operations Director); Andy Taylor (Education Director for Literacy); Clare Brown; and Alia Novak.

Qa Research: Katie Morris (Research Manager); and Rachel Brown (Senior Education Executive).

The Education Endowment Foundation: Rachael Morris (Senior Evaluation Manager); Katie Luxton (Senior Programme Manager); and Fabiola Clemente (Programme Manager).

Executive summary

The project

The Reciprocal Reading programme is a targeted intervention, developed by Fischer Family Trust Education (hereafter FFT), which aims to improve pupils' reading comprehension and, in the longer term, their overall literacy. This intervention is designed for pupils who read words accurately but often struggle to understand the meaning of what they read. It aims to develop pupils' understanding of a text through the application of four strategies—predict, clarify, question, and summarise—used repeatedly on small sections of the text, to deal with comprehension difficulties as they emerge.

The intervention is delivered by trained teachers and teaching assistants for two 20–30-minute sessions per week, for a minimum of 12 weeks, to pupils in Years 5 and 6, identified as having reading comprehension difficulties using FFT's screening tool. At least 12 pupils receive the intervention in groups of four to eight, in addition to normal reading/English lessons. To support scale-up, Reciprocal Reading uses a train-the-trainer model. FFT trainers deliver the initial training and ongoing support to school staff, with a senior trainer overseeing the FFT trainers and monitoring delivery quality.

The trial was conducted in 295 primary schools across England throughout the academic year 2023/2024. This large-scale effectiveness trial used a two-arm clustered randomised controlled trial, complemented by a mixed methods implementation and process evaluation (IPE). The IPE included in-depth case studies in four schools. This combined semi-structured observations of Reciprocal Reading sessions, interviews with school coordinators and teachers/teaching assistants, and focus groups with pupils. It also included interviews with a sample of 20 parents of pupils in the treatment group.

Table 1: Key conclusions

Key conclusions	
1.	Pupils in Reciprocal Reading schools made one month's more progress in reading on average, compared to pupils in other schools. This result has a high security rating.
2.	Among pupils eligible for free school meals (FSM), those in Reciprocal Reading schools made no additional month's progress in reading, on average, compared to pupils in other schools. These results have a lower security than the overall findings because of the smaller number of pupils.
3.	Supplementary analysis revealed that intensity matters. Pupils who received both the recommended intensity and minimum number of sessions (at least 20 sessions over 12 weeks or less, as opposed to more spaced-out delivery) received the equivalent to two months' progress, as compared to the average pupil in the intervention group.
4.	Teachers perceived that training was of a high quality and followed the delivery model closely. High attendance and positive engagement with the training sessions meant that teachers delivered the intervention as intended.
5.	Teacher observations were of a high quality, with teacher attitudes towards the intervention reported as overwhelmingly positive, and teachers reporting high levels of pupil engagement given the interactive nature of the intervention.

EEF security rating

These findings have a high security rating. This was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of settings. The trial was a well-designed, two-armed, randomised controlled trial. The trial was well powered. Relatively few pupils (9%) who started the trial were not included in the final analysis. The pupils in Reciprocal Reading schools were similar to those in the comparison schools in terms of prior attainment. Some control schools delivered parts of the Reciprocal Reading programme, which makes it harder to accurately estimate the size of the impact on the pupils in the trial.

Additional findings

Pupils in the Reciprocal Reading schools made, on average, one month's additional progress in their reading score compared to those in the control group. This is our best estimate, which has a high security rating. As with any study, there

is always some uncertainty around the result: the possible impact of this programme also includes no additional progress and positive effects of up to two months' additional progress. While this analysis is exploratory, it appears that the intervention improved reading accuracy and comprehension at the sentence level but not passage comprehension, which was the developers main target outcome.

There is strong evidence to show that Reciprocal Reading operated as hypothesised by the programme's Theory of Change. Surveys completed by teachers before and after their training show large increases in their self-reported confidence in implementing Reciprocal Reading, knowledge of Reciprocal Reading concepts, and actual increases in knowledge. Interviews with teachers, further support that the training worked as expected and increased their confidence, knowledge of concepts, and overall knowledge.

Teachers believed that several pupil-level factors influenced how effective the programme was for different groups. Only 5% of teachers reported a boost in confidence for pupils with English as an Additional Language (EAL), while 82% of teachers felt Reciprocal Reading was more effective than usual teaching for pupils with stronger prior reading skills. Many staff saw clear benefits for FSM-eligible pupils, highlighting the programme's structured, interactive, small group approach as particularly supportive, though the subgroup analysis suggests that pupils eligible for FSM made similar progress to pupils in the control group. Pupil engagement was consistently high (98%), with the discussion-based format making reading feel more enjoyable. Overall, effectiveness appeared greatest where engagement and prior skills were strong, and weaker where language or home factors created barriers.

Key challenges to implementation included timetabling issues (65% of teacher survey respondents), insufficient staff resources (32%), and lack of physical space (29%). Despite these, most schools managed to find ways around them, often by being flexible or using teaching assistants for delivery.

Targeted reading support of varying kinds was delivered to pupils in the control group. Around 38 schools in the control group used reading interventions from other organisations including PiXL (Partners in Excellence), FFT (Tutoring with the Lightning Squad), and Reading Plus, which may help to explain the smaller impact found in this trial.

The impact on the New Group Reading Test overall reading attainment score in this trial was approximately half the size of the impact found in the previous efficacy trial (O'Hare *et al.,* 2019). This is to be expected. It is common for programmes to be less effective on average when delivered at scale.

Cost

The average cost for one school to implement Reciprocal Reading for three consecutive years is £2,759 or approximately £77 per pupil per year, assuming the intervention is delivered to 12 pupils each year (total of 36 pupils).

Impact

Table 2: Summary of impact on primary outcome

Outcome / group	Effect size (95% confidence interval)	Estimated months' progress	The EEF security rating	No. of pupils	P-value	EEF cost rating
Reading attainment	0.06 (-0.14, 0.12)	+1		3,878	0.117	£ £ £ £ £
Reading attainment, pupils eligible for FSM	0.04 (-0.06, 0.13)	0	N/A	1,525	0.465	N/A

N/A = not applicable.

Introduction

Background

Policy context

Key Stage 1 reading attainment has reduced substantially in England since 2019. In the 2022/2023 academic year, 68% of pupils met the expected standard in reading, down from 75% in 2019 (DfE, 2022). During the same period, the reading attainment gap between pupils from disadvantaged backgrounds[1] and pupils not known to be disadvantaged widened (DfE, 2022). To help solve this problem, the DfE reendowed the Education Endowment Foundation (EEF) to produce high-quality research on what works to improve attainment in schools, including literacy. As part of this vision, the DfE's Accelerator Fund supports the EEF research, including this effectiveness trial of the Reciprocal Reading targeted programme. One aim of the fund is to scale evaluations of educational interventions that have shown promise in previous trials. To fully evaluate whether these interventions are effective gap closers, the EEF is focusing on scaling projects in the Education Investment Areas.[2]

Existing evidence

Until recently, there was a lack of experimental evidence rigorously evaluating the impact of Reciprocal Reading as a clearly defined intervention in the UK context. After being developed in New Zealand in the 1980s, the approach was evaluated primarily in the United States (US) and Australia. Rigorous quantitative evidence was collated to show the impact of reciprocal teaching, which has many parallels to Reciprocal Reading as an approach to improve educational attainment. In 1994, the available evidence from the international literature was integrated in a meta-analysis of 16 quantitative studies that computed an average effect size of +0.32 on reading comprehension (using ad hoc tests rather than standardised tests) (Rosenshine and Meister, 1994). However, while this meta-analysis shows that reciprocal approaches showed promise for improving educational attainment, they cannot be used as surrogates for studies evaluating Reciprocal Reading as a defined intervention.

In 2014, the EEF conducted the first UK trial of the Reciprocal Reading programme. It was a small-scale efficacy trial with 41 secondary schools, which yielded inconclusive results (Crawford and Skipp, 2014). Despite finding a small positive effect on Year 7 pupils, issues with testing, high attrition, and sample imbalance meant causality could not be established (Crawford and Skipp, 2014). In 2019, the EEF and Queen's University Belfast conducted an efficacy trial to test Reciprocal Reading for Key Stage 2 pupils in 98 English primary schools. This study investigated the effects of an intervention on the reading skills and comprehension of Year 4 pupils as a whole class, and as a targeted intervention delivered to a smaller group of 12 Year 5 and Year 6 pupils (aged 9–12) who were identified as good at decoding but poor at comprehension. Although the results did not show evidence of impact,[3] in months of progress, on reading skills, and comprehension for the whole class, the intervention showed promise for the targeted model. The targeted intervention improved overall reading scores by +0.14 standard deviations (SDs), as measured by the New Group Reading Test (NGRT) standardised test, equivalent to two additional months' progress (O'Hare *et al.,* 2019).

Between 2021 and 2022, two more UK-based randomised controlled trials evaluated the impact of the targeted Reciprocal Reading model on older pupils aged 11–12 years. The first included 315 secondary school pupils in 14 schools from areas of high social-economic deprivation in England (Thurston *et al.,* 2020). Results were consistent with the 2019 efficacy trial, estimating a significant effect of +0.19 SDs on overall reading.[4] Despite showing that these effect sizes and directions could

[1] The Department for Education (DfE) defines a disadvantaged pupil as a pupil who has been eligible for free school meals (FSM) over the last six years or has spent at least one day in care with the local authority.
[2] The full list of Education Investment Areas can be found in the United Kingdom (UK) Government website at: www.gov.uk/government/publications/education-investment-areas/education-investment-areas
[3] This was possibly due to three factors: i) a lack of targeting of the pupils who benefit most according to the theory; ii) some pupils struggled to access the programme, possibly due to their age (which was one to two years younger than those in the targeted intervention; and iii) the groups were too large for effective facilitation of the intervention (O'Hare *et al.,* 2019).
[4] Measured with the NGRT tests, adapted for secondary school pupils.

be sustained across different settings and adaptations, the authors highlighted the need for a larger scale trial to see if effects generalise to a broader population (Thurston *et al.,* 2020). The second studied a small sample of 800 pupils over 20 English schools and found no significant differences between control and intervention groups (Cockerill *et al.,* 2025).

Building on this evidence, a 2021 mixed methods process evaluation conducted across 35 schools in an area of high deprivation suggested the intervention could be rolled out with high fidelity across Key Stages 1 and 3 (Cockerill *et al.,* 2022). The study claims to provide evidence that the intervention can be implemented in a variety of settings and in regions of high deprivation. However, due to the relatively small sample of schools in the study, these claims need to be assessed with caution.

Rationale for this evaluation and its design

A large-scale trial is needed to determine with confidence whether the targeted version of Reciprocal Reading can be effective at scale across England. The EEF therefore, commissioned the Behavioural Insights Team (BIT) to lead a large-scale, independent evaluation. We ran a two-arm clustered randomised controlled trial in 295 primary schools across England, alongside a mixed methods implementation and process evaluation (IPE). The EEF previous efficacy trial was conducted in ideal conditions with close support from the developers. In this effectiveness trial, the aim was to test the targeted intervention at scale with minimal interaction from the developers to observe the impact in real-world conditions. This involves a large number of settings, as well as following the development and implementation of a train-the-trainer model. This new model is what was evaluated in the present trial. To allow for a clear comparison of results between this study and the efficacy trial that preceded it, we replicated aspects of the efficacy trial design, particularly the outcome measures.

Intervention

This intervention description uses an adapted version of the Template for Intervention Description and Replication (TIDieR) framework (Hoffman *et al.,* 2014) for its structure. A description was developed by the Reciprocal Reading team at Fischer Family Trust Education (hereafter FFT) and researchers at Queen's University Belfast for the efficacy trial of Reciprocal Reading (O'Hare *et al.,* 2019, p. 11). This new version has been created to add detail and to reflect key changes to the programme since the efficacy trial.

Why does the programme exist?

The programme aims to improve pupils' reading comprehension and, in the longer term, their overall literacy. It addresses this goal by teaching children to use four key reading comprehension strategies: predict; clarify; question; and summarise. These strategies help children to understand the texts that they are reading in the intervention and also provide a framework for them to apply to texts (both fiction and non-fiction) that they read in the future.

What does the programme entail?

The programme has two key components:

1. A reading programme for pupils.
2. Training and support for school coordinators, teachers, and teaching assistants who deliver the reading programme.

These are described in turn below.

Reading programme for pupils

The programme for pupils involves a series of reading sessions for small groups (approx. six pupils per group, with a minimum of four pupils and a maximum of eight pupils). In each session, pupils read part of a text at a time, individually, in pairs, or as a whole group. Pupils are supported to read the text using four reading strategies:

1. **Predict.** Pupils make a prediction about the text; for example, what will happen to a character in a story.
2. **Clarify.** Pupils identify words or phrases that they do not fully understand and work out their meaning through discussing the unfamiliar words in context.
3. **Question.** Pupils ask questions of the text to deepen their understanding; for example, why a character is behaving in a certain way, and are encouraged to answer the questions through group discussion.
4. **Summarise.** Either the teacher/teaching assistant leading the session or a pupil will summarise what has been read and other pupils will be invited to add information.

The aim is for pupils to be taken through this cycle of strategies three or four times per session. The texts that pupils read in the sessions can come from the anthology of short stories provided in the programme materials or be chosen by the teacher/teaching assistant following the guidance in the resource materials.

Training and support for school coordinators and teachers/teaching assistants

The teachers and teaching assistants who coordinate and deliver the reading programme are given the following training and support:

1. **Programme briefing.** An initial briefing for the teacher responsible for coordinating the intervention in their school that includes guidance on how to choose the pupils for the intervention.
2. **Training.** Two days of training for teachers and teaching assistants responsible for delivering the programme with pupils.
    a. Day 1 covers the principles behind the intervention, the programme process, resources for delivery, planning for implementation, the opportunity for participants to experience a Reciprocal Reading lesson themselves and watch recordings of its use in school, and guidance and tools for quality assurance (QA).
    b. Day 2 focuses on extending practice so that it continues to meet the needs of more experienced and improving readers by challenging them further.
3. **Online support meetings.** To provide guidance on implementing and sustaining the intervention.
4. **Ad hoc support.** An email and telephone helpline, which school coordinators can contact for advice on implementation.

Who are the participants?

Participants need to meet the following criteria to be eligible for the Reciprocal Reading programme: baseline demographics; and baseline reading proficiency.

Baseline demographics

Aged 9 to 11 (primary school Years 5 and 6).

Baseline reading proficiency

The intervention is designed for pupils who have good decoding skills but poor comprehension skills. These are pupils who struggle to make sense of the texts they read, particularly struggling to understand anything, which is not explicitly stated in the text. The teacher or teaching assistant will identify these pupils based on their perceptions of the pupil's reading proficiency and on a checklist of characteristics provided by FFT as part of the introductory webinar, which includes a guidance document on screening pupils.

Teachers and teaching assistants are asked to supplement their perceptions with existing school assessments, looking for results showing high scores on reading accuracy but significantly lower on reading comprehension, and using the Simple View of the reading model to reflect upon the information gathered (Gough and Tunmer, 1986). FFT provides a guidance document on how to make an informed selection but is not compulsory for schools to use.

The rationale for these eligibility criteria is that there are already explicit interventions for children whose decoding skills are weak, and children who are average or good readers can access whole-class teaching easily and read independently.

Children who can decode but struggle to understand are a hidden group who may be functioning just below age-related expectations, not reading for pleasure, and unlikely to be receiving targeted reading support. Intervention developers hypothesise that pupils who can decode but struggle to comprehend texts may find it harder to draw pleasure from reading and thus, may read less—which in turn may affect their reading attainment (Clark and De Zoya, 2011). This creates a cyclical issue, which links poor comprehension to less engagement with reading, leading to lower reading frequency, resulting in lower reading attainment. By identifying these readers, teaching staff are supported to think about their needs and begin to address their weaknesses, improving their ability to read with understanding and to read for pleasure.

What materials are needed?

FFT's Reciprocal Reading trainers provide the teachers and teaching assistants delivering the sessions and school coordinators with a resource pack that will help them deliver the intervention in school. The pack contains the following:

1. A handbook that contains:
   a. A summary of the evidence behind the intervention.
   b. Detailed descriptions of the four strategies.
   c. Formats and templates for planning sessions.
   d. Examples of session plans.
   e. Advice on how to choose appropriate texts for pupils to read.
   f. Support materials for Training Days 1 and 2.
   g. Notes on the core components of the training, cross-referenced with the materials given to schools.
   h. Advice on how to conduct follow-up support meetings with schools, including frequently asked questions.
2. Advice on how to choose the pupils for the programme.
3. A manual on programme implementation, outlining the main expectations of the programme for the school when delivering the programme, the delivery timeline, and the roles and responsibilities for the school staff involved in the programme.
4. An anthology of short stories to use in the sessions.
5. Dictionaries for pupil use.

In addition to the resource pack, schools are provided with access to a website with a video material of lessons and additional planning examples.

Who delivers the programme?

Table 3 gives an overview of the roles involved in delivery of the intervention.

Table 3: Role descriptors for the project

Role title	Role description	Organisation
School coordinator	A member of teaching staff (ideally a senior leader): • Manages the intervention in a school. • Supports the selection of teacher/teaching assistant who will deliver the sessions. • Support the selections of target pupils, which will be led by their current teacher or another member of the senior leadership team (SLT). • Monitors attendance to ensure minimum dosage is delivered. • Offers ad hoc support to teachers/teaching assistants on session delivery. • Advocates for the intervention with senior leadership and governing boards. • Encourages implementation of Reciprocal Reading practices after the intervention has ended.	School
Session lead	A teacher or teaching assistant: • Delivers the programme reading sessions with pupils. • Keeps attendance register.	School

Role title	Role description	Organisation
FFT trainer	• Delivers the training and support programme for school coordinators and teachers/teaching assistants.	FFT
FFT school support team	• Provides ad hoc email and telephone advice to school coordinators and teachers/teaching assistants. This would then be triaged to the appropriate support resource.	FFT
FFT senior trainer	• Trains the FFT trainers. • Facilitates the community of practice for FFT trainers. • Monitors implementation data and the quality of delivery in schools.	FFT

How is the programme delivered?

The programme for pupils is delivered in curricular time. Each school will decide how to integrate it into the timetable, but FFT recommends avoiding teaching them in place of curricular English lessons or individual reading lessons. In fact, the intervention should not replace other curricular lessons or activities generally even from other subjects.

The training for teachers and teaching assistants is delivered in addition to schools' standard programme of in-service training (INSET) days.

When, where, and how much?

Table 4 summarises the location, timing, and duration of all programme activities.

Table 4: Time, location, and duration of intervention activities

Activity	When	Where	Duration
Reading programme for pupils			
Reading sessions	Two sessions per week, for 12 weeks, during Autumn Term 2023 and Spring Term 2024	In school, in a reasonably quiet and uninterrupted working environment	20–30 mins per session
Training and support for teachers and teaching assistants			
Programme briefing	April 2023 to September 2023	Online	90 mins
Training Day 1	October 2023 and November 2023	Out-of-school training facility	One day
Training Day 2	February 2024 and March 2024	Out-of-school training facility	One day
Support meetings	One between Training Days 1 and 2. One after Training Day 2 (end of Spring Term / beginning of Summer Term)	Online	Two x 60 mins
Ad hoc support	When needed during delivery period	Email or telephone	As needed

Tailoring

Table 5 indicates which of the programme activities are core (i.e. should be delivered always) and what kind of adaptations are expected and acceptable.

Table 5: Core components and acceptable adaptations

Activity	What's core?	Acceptable adaptations
Reading programme for pupils		
Reading sessions	Core delivery: • Two weekly 20-min sessions over 12 weeks. • At least four pupils per group (with a maximum of eight pupils). • Sessions follow the four-strategies structure, repeated three or four times per session. • Must read an appropriate text as defined in the intervention guide. • Dictionaries need to be available to support word clarification. • Working environment must be reasonably quiet and uninterrupted. Core management: • Selecting pupils according to criteria. • Monitoring of session delivery and attendance. • Ad hoc support from school coordinators to teachers/teaching assistants on session delivery. • School coordinator must support timetabling. • Minimum of half-termly meetings between school coordinators and teachers/teaching assistants.	Delivery adaptations: • Delivering more than the minimum dosage (in terms of session length and number). • Pupil groups can either be separated by year group or be mixed year group (within pupil eligibility criteria). • Some pupils may leave/join the group after the start, always following pupil selection guidance for newcomers to avoid altering the small group dynamic. • Some groups might read texts not in the anthology, using text selection guidelines provided. • The extent to which ad hoc programme support is drawn upon. • The extent to which teachers and teaching assistants use the four reading strategies in other subjects (embeddedness). Management adaptations: • Level and type of QA and support from school coordinators • Whether or not the school coordinator does intervention profile raising work with the SLT and governing boards.
Training and support for teachers and teaching assistants		
Programme briefing	• All of session. • School coordinator and sessions lead must attend.	• Replacement teachers/teaching assistants receive online training from FFT.
Training Day 1	• All day.	• Replacement teachers/teaching assistants receive online training from FFT.
Training Day 2	• All day.	• Replacement teachers/teaching assistants receive online training from FFT.
Support meetings	• Both meetings.	• Can last between 60–90 mins depending on need. • Meetings can happen in person at the school if the trainer lives nearby. • Support meeting can be for the school coordinator alone or with teacher/teaching assistant delivery staff.
Ad hoc support	• Must be available to all schools.	• Optional for schools to use.

Tailoring delivery to scale: Training the trainers

To scale the intervention delivery, FFT trained a new cohort of trainers. All trainers were members of the education team at FFT, they are all previous teachers or SLTs in schools with a reading/literacy specialism. Some had taught it in school themselves but not trained on it. While this was not a key component of the intervention, it was a key activity to tailor the intervention to this effectiveness trial. Before delivering training in schools, trainers completed the following development activities:

1. **Initial training.** Three and a half days training on the principles of Reciprocal Reading, and how to deliver the training and support programme for teachers and teaching assistants.
2. **Observations.** At least one observation of an experienced trainer delivering the two days of training for teachers and teaching assistants.

3. **Co-training.** Delivering at least one full training programme for teachers and teaching assistants alongside an experienced trainer.
4. **Independent delivery of training pre-trial.** Delivering the full training programme for teachers and teaching assistants at least once before starting training trial schools.
5. **Community of practice.** Facilitated discussions to share practice between trainers during the delivery period.

The trainers were trained by experienced staff from FFT, 'FFT senior trainer'. Two cohorts of trainers were trained for the effectiveness trial. Cohort 1 (three trainers) received all five elements of the training and support described above before the trial began. Cohort 2 (two trainers) received an equivalent training that started before the trial began and continued in parallel to the trial. Trainers could not start training teachers and teaching assistants on Reciprocal Reading until they had completed their own training. Important to note, in the efficacy trial, the training was delivered by a small team of highly experienced trainers (who now act as FFT senior trainers), so this represents an important change in the delivery approach.[5]

Managing quality

FFT put the following strategies in place to support effective implementation:

- A clear set of written programme materials, including a manual, a set of suggested session plans, and anthology of short stories.
- A comprehensive and standardised training and support programme for school coordinators and teachers/teaching assistants.
- Monitoring and support from the FFT senior trainer and the wider FFT education team, including the following QA observations by the FFT senior trainer:
    - At least one observation of each FFT trainer covering Training Day 1 and Day 2 and support meetings.
    - Observations of a sample of schools, inspecting the programme set-up.
- Guidance on pupil selection in the written materials and webinars, and additional advice and support on the choice of pupils available to schools upon request.
- Schools are given an online tool to register attendance and dosage of the Reciprocal Reading sessions. Teachers/teaching assistants will keep registers for all of their sessions. FFT and the school coordinator will have access to these for monitoring.
- A programme of online training to train replacement teachers/teaching assistants in the event of staff turnover during intervention delivery.
- Two support meetings with the trainers, which are part of the core delivery, to provide guidance and discuss implementation in the school.

[5] Cohort 1 trainer training took place before the start of the trial so the train-the-trainer model is not covered in detail in the IPE.

Logic models

Figure 1: Logic model on training for teachers/teaching assistants and the school coordinator

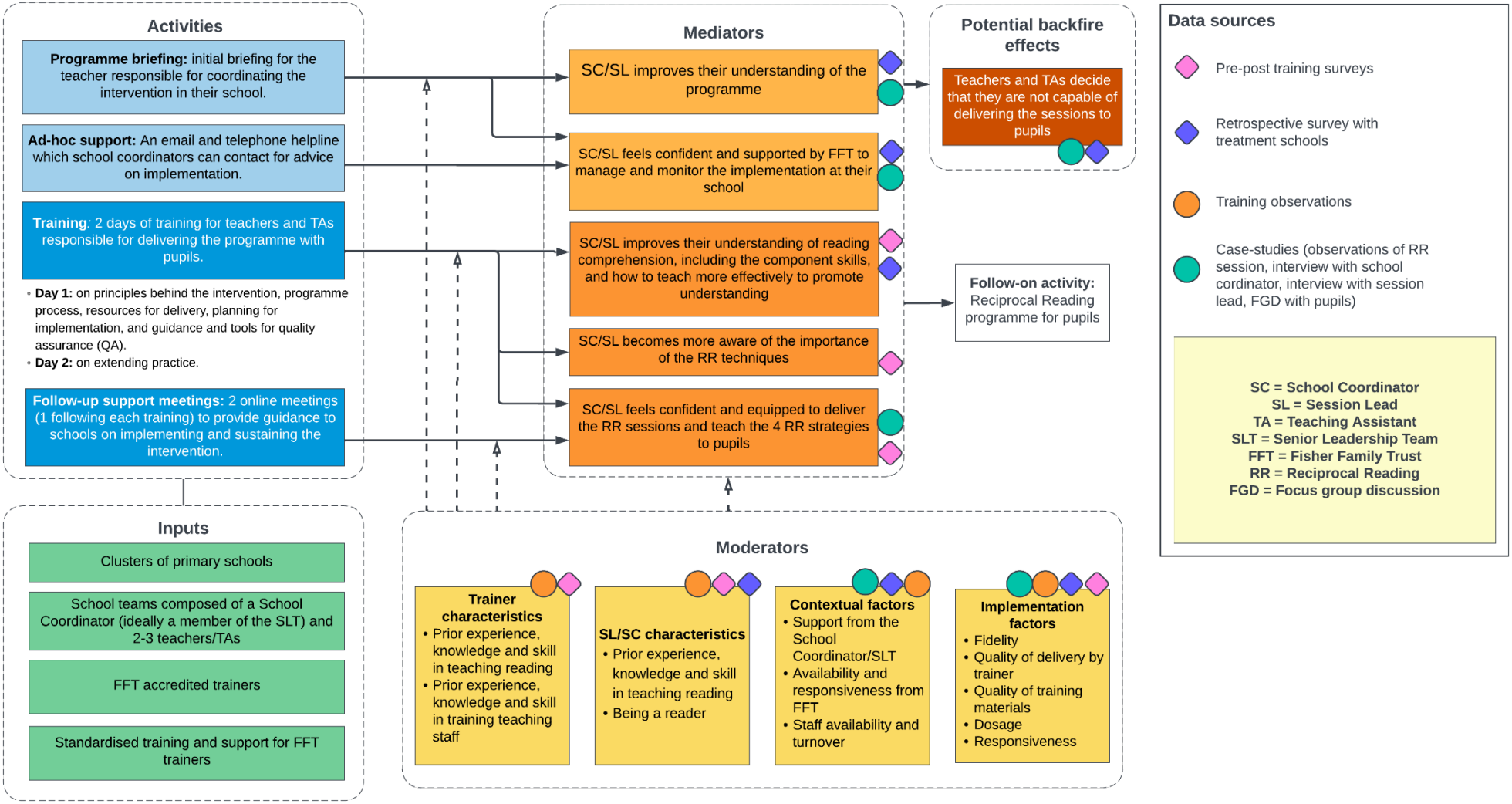
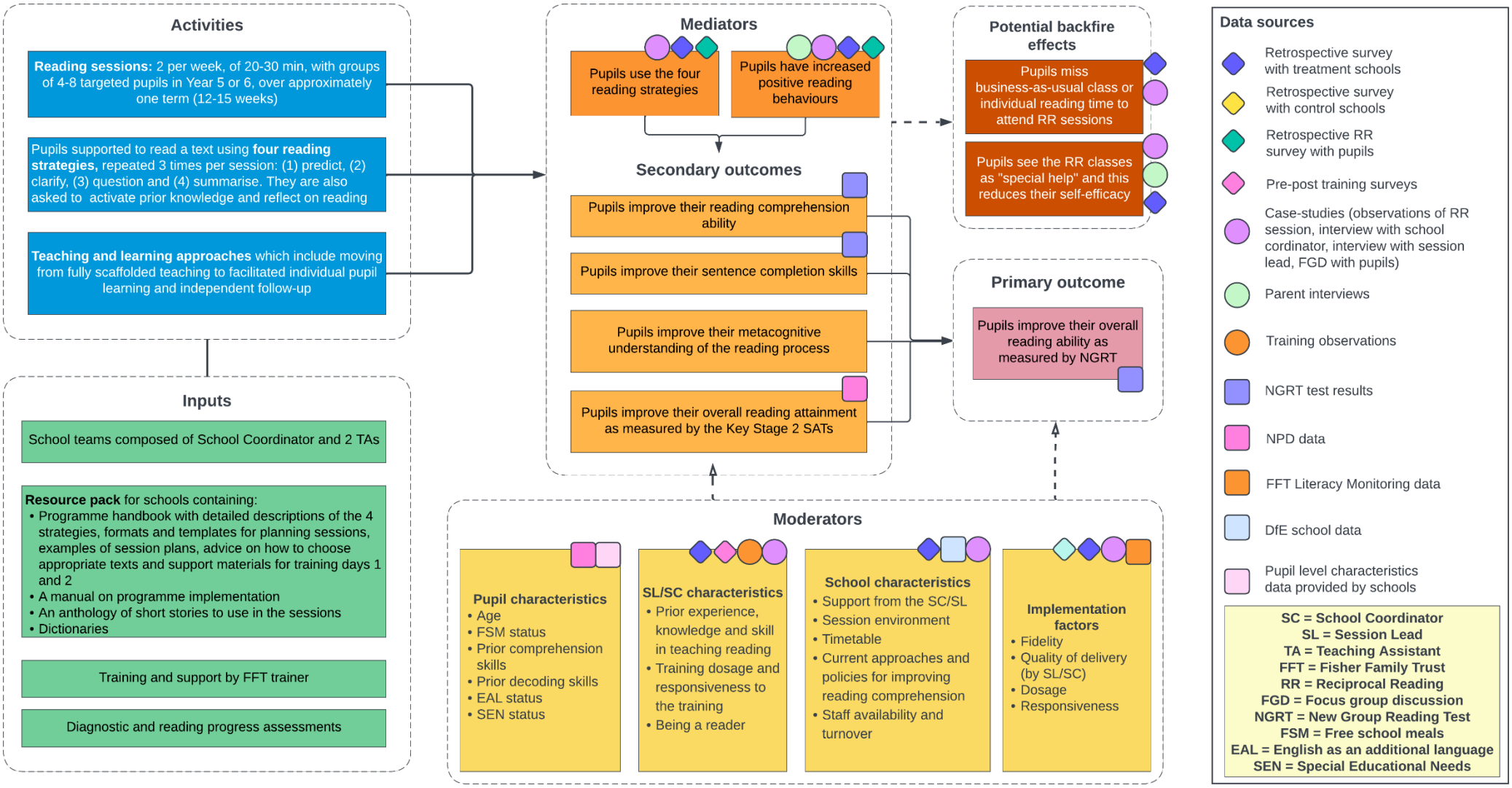


Figure 2: Logic model on Reciprocal Reading pupil outcomes



Evaluation objectives

Impact evaluation

The impact evaluation aimed to address the following primary research question:

1. What is the difference in reading proficiency (as measured by the NGRT overall reading score) of pupils who receive the Reciprocal Reading programme, as compared to pupils in the control group?

There were a further four secondary research questions:

2. What is the difference in reading proficiency (as measured by the NGRT overall reading score) of FSM-eligible pupils who receive the Reciprocal Reading programme, as compared to FSM-eligible pupils in the control group? Is the treatment effect for FSM-eligible pupils different from the effect for the full sample of pupils?
3. What is the difference in reading comprehension (as measured by the NGRT passage comprehension score) of pupils who receive the Reciprocal Reading programme, as compared to pupils in the control group?
4. What is the difference in sentence completion skills (as measured by the NGRT sentence completion score) of pupils who receive the Reciprocal Reading programme, as compared to pupils in the control group?
5. What is the difference in reading attainment (as measured by the Key Stage 2 SAT reading score) of pupils in Year 6 who receive the Reciprocal Reading programme, as compared to pupils in the control group?

IPE

The IPE accompanying the trial aimed to address seven key research questions:

1. To what extent did trainers deliver the training as intended, and how well did they deliver it? (Fidelity/Quality)
2. To what extent did school coordinators and teachers/teaching assistants implement Reciprocal Reading sessions as intended, and how well did they deliver it? (Fidelity/Quality)
3. What were the characteristics of the schools and pupils reached through scaled-up delivery, and how did these characteristics moderate the effects of the intervention? (Reach/Moderators)
4. How much of the intervention did treatment group pupils receive? (Dosage)
5. How did the intervention work for pupils from disadvantaged socio-economic backgrounds? (Moderators/Causal mechanisms)
6. Did the intervention complement or replace existing efforts? What did teachers and teaching assistants perceive to be the distinguishing features of the intervention? (Programme differentiation/Unintended consequences)
7. What reading interventions did selected pupils in the control group receive? And did the pupil selection process seem to influence this? (Adherence, Contamination)

Key links:

- Trial Protocol (Cappellini *et al.*, 2022).
- Statistical Analysis Plan (Torres Blas and Taylor, 2019).

Ethics and trial registration

Ethical approval

Ethical review was conducted by the BIT Internal Ethics Committee, and the project was approved on 14 November 2022. The review was conducted by a non-project member of staff, who blind assessed the research plan and data collection materials to ensure the project had considered all the potential risks and had appropriate procedures in place to mitigate these. In this case, the Ethical Review was conducted by Dr Giulia Tagliaferri, Head of Quantitative Research at BIT.

Following BIT's ethical framework, the project was assigned a medium-risk level. This is due to the following key considerations:

- Low risk in terms of research methods, subject matter, legal exposure, and unknown unknowns.
- Medium risk in terms of participants (some children in a regular setting).
- High risk in terms of nature of data (some personal, individual level, special category data collected for pupils).

Procedures for obtaining agreement to participate in the trial

As part of the recruitment process, schools were sent the following documents to lay out the requirements and terms of the trial.

- **Memorandum of Understanding (MOU) (see Appendix E).** The MOU explained the programme, what was required of schools, data protection, and expected activities.
- **School information sheet.** Provided further information about all topics in the MOU and explained the study design, including the process for randomisation.

We also shared with schools a Parent Information Sheet and Consent Withdrawal form, which explained the programme and trial to parents and gave them the opportunity to withdraw their children from the trial.

These documents provided clear explanations of the programme and evaluation and gave schools the chance to make an informed decision about participating in the trial. These were designed collaboratively by the trial partners following the EEF specific guidance.

Trial registration

The trial was registered on Open Science Framework (OSF). It was registered by the BIT evaluation manager on 07 August 2024. The evaluation manager will be responsible for updating the registry with the trial results after the report is published in 2026.

The trial registry DOI is: https://doi.org/10.17605/OSF.IO/8RHFD.

Data protection

All pupil data has been processed in accordance with the General Data Protection Regulation (GDPR) (2016) and Data Protection Act (2018).

Data security

BIT has robust approaches to protecting personal data that comply with the Data Protection Act (2018) and GDPR (2016). We conduct all projects with a privacy by design approach to protect and maintain the privacy and security of research participants' data. All staff are trained on GDPR compliance and BIT's data protection policies. BIT is registered with the UK Information Commissioner's Office (ICO) under the terms of the Data Protection Act (2018) and has a full-time data protection officer. BIT has obtained Cyber Essentials Plus certification and ISO 27001 certification (the international standard for information security). BIT takes steps to protect participants' personal information and prevent unauthorised access, alteration, loss, or disclosure. We train those with access to personal data, provide access only on a need-to-know basis, and remove access when it is no longer needed. Devices are encrypted, confidential data is kept securely, and we have procedures in place to deal with data breaches and notify participants and regulators if necessary.

Data retention

The data collected during the Reciprocal Reading Evaluation Research Project will only be kept for as long as necessary and will be deleted securely when it is no longer needed. The appropriate retention period is determined by considering factors such as the amount and sensitivity of the data, potential risk of harm, and legal requirements. The anticipated deletion date for participants' personal data is three months after the project's completion. However, names and work contact details of school staff may be kept for future similar projects. Research consent forms with personal information may be kept for a number of years for legal and statutory requirements and to meet the research funder's requirements. Personal data is needed to evaluate the impact of Reciprocal Reading in schools.

At the end of the project, a dataset will be archived with the EEF, following their guidelines. We will preserve anonymity and confidentiality in all publications, and no school or individual will be named or identifiable.

Data rights

To exercise their data rights, research subjects can contact the BIT's data protection officer. The data protection officer at BIT can be contacted via email at dpo@bi.team. It should be noted that the extent to which these rights apply to research may vary and may be restricted in some circumstances. Typically, there is no fee required to access personal data or exercise rights, but a reasonable fee may be charged if the request is unfounded, excessive, or repetitive. If the request is deemed as such, BIT may refuse to comply with the request. BIT may request specific information to confirm the identity of the requestor and ensure that the personal data is not disclosed to an unauthorised party. In some cases, additional information may be requested to expedite the response.

BIT aims to respond to legitimate requests within a month, but if the request is complex or multiple requests have been made, it may take longer. In such a scenario, the requestor will be notified and updated on the status. It should also be noted that BIT can only comply with requests to exercise rights for personal information that directly identifies the requestor. If the information is pseudonymised or has been irreversibly anonymised and is part of the research data set, BIT will not be able to comply with the request.

Data collected

The following data was collected from research participants to complete the trial.

Pupil data

We collected personal data from pupils involved in the research including:

- Unique Pupil Number (UPN, identifies the pupil in the National Pupil Database [NPD]);
- first name;
- last name;
- date of birth;
- school name;
- year group;
- NGRT pre-intervention and post-intervention scores;
- Key Stage 2 SATs reading scores (for pupils who were in Year 6 at the start of the trial);
- gender;
- FSM status;

- ethnicity;[a]
- whether the pupil speaks English as an Additional Language (EAL);
- Special Educational Needs and Disabilities (SEND) status;[a]
- SEND type—type of special educational need;[a]
- focus group discussion data—qualitative data gathered in focus groups on attitudes to the Reciprocal Reading intervention; and
- Reciprocal Reading session attendance data—data on the number and length of Reciprocal Reading interventions attended.

[a] Some of this data—ethnicity and data about health conditions—constitutes 'special category data' under data protection laws, and additional protections applied to our collection and use of this data. This information is vital for our analysis, to assess the reach and possible differential effects of the programme. Reporting on these fields for the purpose of our research will be in an aggregated format only.

Pupils' parent data

We collected personal data from parents of the pupils involved in the research including:

- first name;
- last name;
- email address;
- telephone number;
- data on socio-economic background (FSM status) of their children; and
- interview data—qualitative data on attitudes to reading and self-reported reading behaviours in the home.

School staff data

We collected personal data from school staff involved in the research including:

- first name;
- last name;
- work email address;
- work telephone number;
- survey data—quantitative and qualitative data on attitudes to the Reciprocal Reading intervention; and
- interview data—qualitative data on attitudes to the Reciprocal Reading intervention.

BIT also used pupils' UPNs to access the following data using the DfE's NPD:

- Key Stage 2 SATs reading scores;
- FSM status;
- gender;

- ethnicity;
- EAL status—whether an EAL pupil;
- SEND status; and
- SEND type—type of special educational need.

BIT shared participants' data with the following organisations:

- **Qa Research.** An independent provider contracted to administer the NGRT in partner schools.
- **GL Assessments.** An independent provider which provided the platform to run the NGRT tests.
- **Schools.** Schools participating in the research, which will receive pupil NGRT scores at the end of the study.[a]
- **McGowan.** BIT's transcription provider.
- **SmartSurvey.** Online platform to administer surveys.
- **FFT.** Will receive pupil NGRT scores at the end of the study.
- **DfE.** NPD team.
- **The Office for National Statistics (ONS) Secure Research Service (SRS).** We used the ONS SRS to analyse the trial data securely when matched to NPD data.
- At the end of the evaluation, the pupil data will be shared with the EEF and FFT Education (the EEF's data processor for their archive). All the EEF trial data is stored in the EEF data archive, held within the ONS SRS. The archive does not contain direct identifiers like pupil name, contact details, and month and year of birth, but does hold a Pupil Matching Reference (PMR). The PMR is used for further matching to the NPD and other administrative datasets that may be required as part of subsequent research. We will not use pupil names or school names in any report arising from the research. For information on how the EEF will use and protect participants' data, please see their data protection statement regarding the EEF evaluations.

[a] Schools that want to receive individual-level NGRT results for their pupils will have to sign a Data Sharing Agreement with BIT.

Data processing roles

Data processing roles during the evaluation up to the point of data being deleted from all locations by the evaluator and/or delivery team:

- **BIT.** Independent data controller.
- **FFT.** Independent data controller.
- **Participating schools.** Controllers.
- **DfE.** Controller.
- **The EEF[a]:** Processor.
- **Qa Research.** Sub-processor.
- **GL Assessments.** Sub-processor.
- **SmartSurvey.** Sub-processor.

- **McGowan.** Sub-processor.

[a] The EEF becomes a data controller for the datasets archived after the trial, once internal quality checks have been successfully completed by the archive manager.

Legal basis for processing data

For all information collected, BIT is relying on the lawful basis of legitimate interest.

BIT's lawful basis for processing personal data is based on legitimate interests according to Article 6(1)(f) of the GDPR (GDPR, 2016). The processing is necessary for conducting an evaluation of the Reciprocal Reading programme commissioned by the EEF, which aligns with BIT's business aims of delivering social impact through research and evaluation.

The processing of personal data from pupils, school staff, and parents is required to understand the effects of the programme, as well as to arrange observations and interviews. The processing of interview/focus group data cannot be made anonymous due to unavoidable information that may be revealed during the interviews.

The processing of pupil demographic data, participation data, and outcome data is necessary to ensure correct assessments and to match data for analysis. Anonymous data cannot be used for this purpose.

For special category data, we also rely on scientific research purposes as a lawful basis.

The processing of personal data is necessary for scientific research purposes with the aim of serving the public interest according to Article 9(2)(j) (GDPR, 2016). BIT has implemented appropriate measures to protect the rights and interests of the data subjects, in accordance with relevant laws. The collected data was limited to what is required for the research, and any direct identifying information, such as names or contact details, has been removed where feasible. The processing is not expected to cause harm or distress to the participants and is not intended to be used for making decisions regarding specific individuals.

More information can be found in BIT's Privacy Notice for Reciprocal Reading.

Project team

The intervention implementation was led by the following team at FFT:

- Katie Kielty, Education Product Manager.
- Laura James, Customer Operations Director.
- Andy Taylor, Education Director, Literacy.
- Clare Brown.
- Alia Novak.

The evaluation was designed and delivered by the following staff at BIT:

- Neus Torres Blas, Evaluation Manager and Quantitative Lead (during project implementation and reporting).
- Chiara Cappellini, Evaluation Manager and Qualitative Lead (during project set-up).
- Emma Forsyth, Qualitative Lead.

- Andrés Cueto, Qualitative Research Support.
- Emma Leith, Qualitative Research Support.
- Dr Patrick Taylor, Evaluation Director.

The following BIT researchers provided feedback and QA:

- Dr Laure Bokobza.
- Dr Bobby Stuijfzand.
- Dr Giulia Tagliaferri.

NGRT data collection was managed by the following team at Qa Research:

- Katie Morris, Research Manager.
- Rachel Brown, Senior Education Executive.

Methods

Trial design

Table 6: Trial design

Trial design, including number of arms		Two-arm, cluster randomised controlled trial
Unit of randomisation		School
Stratification variable (s) (if applicable)		Batched randomisation
Primary outcome	Variable	Reading proficiency
	Measure (instrument, scale, source)	NGRT overall reading score, 0–500. GL Assessment (1)
Secondary outcome(s)	Variable(s)	Reading comprehension (2) Sentence completion skill (3) Reading attainment (4)
	Measure(s) (instrument, scale, source)	NGRT passage comprehension score, 0–500, GL Assessment (2) NGRT sentence completion score, 0–500, GL Assessment (3) Key Stage 2 SAT reading score (scaled), 80–120, NPD (4)
Baseline for primary outcome	Variable	Reading proficiency (1)
	Measure (instrument, scale, source)	NGRT overall reading score, 0–500, GL Assessment (1)
Baseline for secondary outcome(s)	Variable	Reading comprehension (2) Sentence completion skill (3)
	Measure (instrument, scale, source)	NGRT passage comprehension score, 0–500, GL Assessment (2) NGRT sentence completion score, 0–500, GL Assessment (3)

This effectiveness trial was a two-arm cluster randomised controlled trial with randomisation at the school level. The two arms were:

1. **Treatment arm.** A treatment arm of schools in which nominated teachers receive Reciprocal Reading training and deliver Reciprocal Reading programme lessons to a group of targeted pupils in Year 5 and Year 6. These lessons were delivered in addition to the school curriculum. These schools received a subsidy from the EEF for the programme fee.
2. **Control arm.** A control arm of business as usual, in which schools continued as they otherwise would have, with no changes in their curricular activities. Control schools were not asked to refrain from any usual support they would provide the targeted pupils. They received a financial incentive of £1,000 to participate in the trial, which was paid after completion of the post-intervention test.

School-level randomisation was chosen to minimise the risk of spillovers between the control and treatment arms, as opposed to pupil- or class-level randomisation. Control schools did not participate in the training for teachers and the risk of communication between treatment and control schools was low.

The impact evaluation estimated the effect of the Reciprocal Reading programme on the overall reading proficiency of Key Stage 2 pupils. This was measured by the overall test score of the NGRT, our primary outcome. Additionally, the evaluation estimated the impact of the programme on three secondary outcomes: pupils' sentence completion skills; pupil's reading comprehension; and reading attainment in Key Stage 2 SATs for Year 6 pupils. The NGRT sentence completion score was used to measure reading accuracy and basic comprehension, while Key Stage 2 reading attainment was measured by the English reading test results of the Key Stage 2 SATs.

Participant selection

Pupil eligibility

The study participants were pupils in Years 5 and 6 (Key Stage 2), aged between 9 and 11 years old, that fulfilled the eligibility criteria in the recruited schools and were nominated by teachers to participate in the trial.

Selected pupils should have had relatively good decoding skills but poor comprehension skills. They were identified by their teacher, with the support of the school coordinator, using guidance and materials provided by FFT when the school signed up for the trial.

The guidance consisted of two sets of eligibility criteria: i) the reading skills that must be held by a pupil to be considered as having good decoding skills; and ii) a list of difficulties that may be experienced by a pupil struggling with reading comprehension. Teachers were instructed to compare each pupil against the two and identify those pupils that matched the criteria.

Teachers were asked to base their assessment on their own subjective judgement but were encouraged to supplement this with existing school assessment data. These assessments should have shown good scores on reading accuracy and lower scores on comprehension tasks for selected pupils. Each school selected two groups of ideally six children in Year 5 and Year 6 that fulfilled both sets of criteria (with a minimum of four pupils and a maximum of eight pupils per group).

Children in the control group schools were selected in the same way for the purpose of completing outcome testing, but they did not receive the Reciprocal Reading programme.

School eligibility

To participate in the trial, a school needed to have at least four pupils in Year 5 and at least four pupils in Year 6 that fulfilled the eligibility criteria and could make two delivery groups; smaller schools that could not reach these numbers could not take part in the trial.

Other school requirements to participate in the trial were:

- having a minimum of one full class of Year 5 pupils and Year 6 pupils (because it was unlikely that a mixed year group class would have enough eligible pupils);
- not having been involved in the previous trials of Reciprocal Reading or used FFT Reciprocal Reading; and
- not being involved in other trials similar to Reciprocal Reading, which focus on improving reading skills, reading attainment, or reading comprehension.

Any schools belonging to a multi-academy trust (MAT) had to agree to not share information about the programme or the teaching of reading comprehension with other schools in the same MAT during the period of the trial, to prevent spillovers between treatment and control schools.

School recruitment was managed and carried out by FFT and aimed to have at least 50% of the recruited schools in Education Investment Areas, as defined by the terms of the Accelerator Fund from the DfE. Recruitment efforts tried to prioritise schools with a higher-than-average proportion of FSM-eligible pupils, although this requirement was not prescriptive and was secondary to achieving the desired recruitment goal of 300 schools needed for statistical power. In treatment schools, recruitment also included one school coordinator per school and two other teachers/teaching assistants (likely to be teaching assistants) that would be trained and would deliver the intervention as teachers/teaching assistants. For control schools, FFT recruited one to two members of school staff (senior leadership, teachers, or teaching assistants) to support activities related to the trial.

FFT recruited schools from January 2023 to July 2023, so programme delivery could start in schools at the beginning of the following academic year (October 2023 – November 2023). Once recruited, baseline testing was carried out in schools on the primary outcome with nominated pupils. They were then randomised into treatment and control by BIT.

Since FFT recruited schools until the very end of the academic year in July, some baseline testing had to be carried out at the beginning of the subsequent academic year and randomisation was conducted in two batches. Approximately 80% of the recruited schools were randomised in July 2023, and the other 20% were randomised in September 2023.

Schools wishing to participate in the trial were asked by FFT to sign an MOU before enrolment, agreeing to the required activities for both the intervention and evaluation. This included the requirement for each school to assess their pupils, select a group of candidates, and allow administration of the baseline NGRT assessment before they are randomised into a trial arm. They also had to repeat the NGRT test with those pupils at the end of the trial. Recruited schools also had to agree to share the relevant pupil data with BIT (via FFT) for evaluation purposes.

This recruitment strategy yielded a sample at randomisation of 295 schools and 4,263 pupils.

Outcome measures

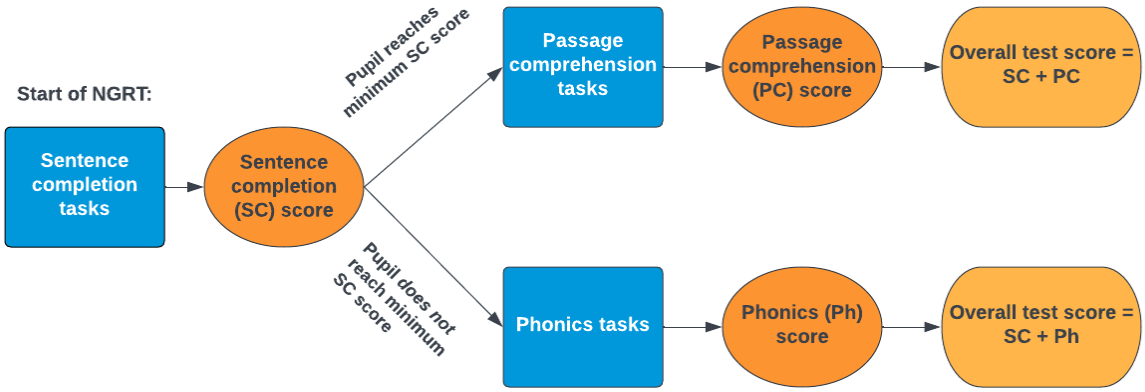
Primary outcome

The trial's primary outcome is the overall score in the digital NGRT by GL Assessment,[6] which measures overall reading proficiency. The NGRT is a standardised assessment that measures skills in sentence completion and reading comprehension and, unlike other national reading assessments like the Key Stage 2 SAT reading test (which is done in May of Year 6), it can be taken by Year 5 and Year 6 pupils immediately after the intervention so can be used for timely measurement. The NGRT overall score was one of the two primary outcomes in the efficacy trial.

The NGRT test was selected because it has a subscale specifically for passage comprehension. This makes it more accurate in measuring the desired outcome than other national English reading assessments and can give more detailed information to teachers about the decoding and passage comprehension skills of a pupil. The digital version of the test (used in this trial) is adaptive, so it responds to the level of the pupil as they take the test. If a pupil does not get a minimum score in the first sentence completion tasks, they will not progress to passage comprehension and will do phonics tasks instead. This means that pupils taking the test receive two subscales that make up the total score: a sentence completion score; and a passage comprehension or phonics score, depending on their performance (see Figure 3). Overall, the reading score is calculated by GL Assessment and available for all pupils but, due to the adaptive progression of the test, reading comprehension scores are not available for those who scored very low on sentence completion tasks, which identify vocabulary gaps. Not being able to achieve the minimum score in sentence completion is likely to mean that a pupil does not have the required decoding skills to access the programme, according to FFT's guidance on pupil selection.

[6] The NGRT is commercial and owned by GL Assessment so we are unable to include a copy of it in the Appendices section of this report.

Figure 3: Diagram of the NGRT online assessment structure



Unlike the efficacy trial, we did not use the NGRT reading comprehension score as a dual primary outcome because it is not available for all pupils taking the test. Comprehension score missingness is not random, as the variable is censored and absent for the lowest achieving pupils, which could bias the estimated effect. We have nevertheless kept the passage comprehension score as a secondary outcome in the impact evaluation, and we also assessed how many pupils were affected by this (see the *NGRT passage comprehension score* section of the secondary analysis).

The NGRT overall test score for the digital version ranges between 0 and 500 points. The test has 47 items, although the total number depends on pupil performance. It has a high reliability with a Cronbach's alpha above 0.9 and was shown to be sufficiently sensitive for the target pupils in the efficacy trial that came before this effectiveness trial (O'Hare *et al.,* 2019). The total score (rather than the standardised score calculated by GL Assessment) is used for the analysis.

The baseline and endline tests were delivered digitally in schools under exam conditions by Qa Research, subcontracted by BIT. Assessors did not know the treatment allocation of the schools they visited. The baseline assessment was done from June 2023 to November 2023 in two batches (June 2023 to July 2023 and September 2023 to November 2023), along with randomisation (we provide more details on the batched randomisation in the corresponding section). The endline was done at the end of the intervention, from March 2024 to July 2024.[7] All pupils took Version A of the test in the baseline and Version B in the endline, to minimise familiarity.[8] GL Assessment's Testwise platform was used to access the online test scores by BIT researchers after the tests were completed. GL Assessment calculated individual pupil scores, which BIT accessed through the online platform. Unique pupil identifications (IDs) were used to set-up initial assessment tests and unique accounts for each pupil. Qa Research, as independent exam invigilators, assisted pupils in using the correct IDs. The testing platform automatically marked and saved the results.

If four or more pupils were absent, Qa Research scheduled a repeat visit for a mop-up assessment. For three or fewer missing results, schools were instructed to self-administer mop-up tests within the following month. Qa Research maintained contact with schools during this period to encourage test completion and provide support for any issues. Around 229 pupil tests (3%) were administered by school staff at baseline across 136 schools, and a further 18 pupil tests (0.2%) were administered by school staff at endline across 12 schools. The school staff administered tests were almost perfectly

[7] This period was needed given the large number of schools in the trial and its national scopes. To maximise response rate, flexibility was given to QA Research when booking test dates with schools. However, to minimise the possibility of bias being introduced by any systematic difference in intervention and control school booking times, they ensured a maximum 60/40% split between intervention and control schools booked per week. As for the treatment schools, FFT monitored their delivery and made sure they had at least implemented the core parts of the programme, which included 12 weeks of delivery and having done the two trainings, before the school could do the endline testing. Treatment schools were also tested in the same order that they had started delivery, with the aim of keeping the dosage similar across schools and to avoid differences in testing being correlated with differences in their implementation.
[8] The NGRT online test has three versions, which are equivalent in terms of difficulty: A; B; and C.

balanced across treatment and control groups. We do not believe that this small proportion of school-administered tests poses a threat to the validity of the trial. Tests are automatically administered and scored by the online platform, providing little opportunity for teachers to influence the results (which they had no motivation to do). The alternative was to not collect any data from these pupils (due to the constrained resources of the data collectors), which was decided to be worse for trial validity due to the reduction in the sample.

Pupils completed the online tests using tablets or computers. According to Qa Research, approximately 20% of the schools experienced IT issues, such as audio malfunctions, unresponsive pages, or upload failures, which necessitated test retakes. Audio issues primarily affected initial instructions, which were mitigated by assessors reading the instructions aloud. While test retakes could introduce risks of test familiarity and score inflation, as well as test anxiety, the NGRT's adaptive algorithm helped minimise grade inflation. The test begins with the same question for all children of the same age and adjusts difficulty based on individual performance. Using anchor points, the algorithm presents questions relevant to each child's ability. A Reading Age Score (RAS) calculated throughout the test, adapts according to the pupil's responses. This adaptivity means that if a pupil familiar with the test performs well initially, the subsequent questions become more challenging, until the pupil reaches a score reflecting their true attainment. This gives the test a high reliability, indicating consistent scores for pupils of the same age and ability. As described below in the 'Statistical analysis' section, we also re-estimate the primary outcome model excluding all pupils from both treatment arms that were flagged by Qa Research as having experienced significant IT issues during endline testing. We compare the results with the main model to assess the extent to which these issues could have affected the main results.

To assess the potential effect of the IT issues on results, we re-estimated the primary outcome model dropping the pupils that were flagged by Qa Research as having experienced a noteworthy IT issue during testing. These results can be found in the 'Additional analyses and robustness checks' section below.

Baseline and endline pupil data were shared with schools upon request after all data collection for the trial was completed in September 2024, provided the school signed a Data Sharing Agreement with BIT.

Secondary outcomes

We also studied the impact of the programme on three secondary outcomes: reading comprehension skills (measured by the NGRT passage comprehension score); sentence completion skills (measured by the NGRT sentence completion score); and reading attainment (measured by the Key Stage 2 SAT reading score).

NGRT passage comprehension score

The NGRT passage comprehension subscale measures performance in reading comprehension tasks. The improvement of reading comprehension is the main target of the Reciprocal Reading programme and therefore, the hypothesised main pathway to increase reading proficiency in the intervention participants (see the logic model in Figure 2). This measure was one of the dual primary outcomes in the efficacy trial so keeping it as a secondary outcome allowed us to compare results between the efficacy and effectiveness trials.

The NGRT passage comprehension subscale is only available for pupils that score a minimum value in the sentence completion tasks of the test and can progress to do passage comprehension exercises. For this reason, we did not use this measure as a primary outcome, even though it is a direct assessment of the target of the intervention.

The NGRT passage comprehension subscale was obtained together with the NGRT overall score, at baseline and endline. The passage comprehension part of the test has 27 items, although the total number of items depends on pupil performance. This is because of the adaptive nature of the test: it presents the next question to pupils automatically based on the pupils' performance as they complete them. This way, higher attaining pupils can be challenged while weaker readers also remain engaged with accessible questions (GL Assessment, 2025a).

NGRT sentence completion score

The NGRT sentence completion subscale measures reading accuracy and basic comprehension through performance on the sentence completion tasks of the test. Although the main focus of the intervention is reading comprehension, the Reciprocal Reading programme also improves overall reading proficiency by consolidating sentence completion skills (see the logic model in Figure 2). We included it as a secondary outcome to test this causal pathway.

The NGRT sentence completion subscale was obtained together with the NGRT overall score and the passage comprehension subscale, at baseline and endline. The sentence completion part of the test has 20 items, although the total number of items depends on pupil performance.

Key Stage 2 SAT reading score

Key Stage 2 SAT reading score was measured using the KS2_READSCORE, obtained from the NPD Key Stage 2 dataset. This is the scaled score on the English reading test, and can range between 59 and 120, where 100 indicates the pupil met the expected standard of the test.[9] This had to be used instead of KS2_KS2READSCORE (the variable recommended in the Trial Protocol; Cappellini *et al.,* 2022) because it was not available, and the differences between the two are minimal. Pupils were linked to the NPD data using their full name, date of birth, and school identifier, which researchers sent to the DfE for linkage to their PMR numbers.

All observations that did not have a numeric value in KS2_READSCORE were considered as missing, regardless of the letter classification by the DfE. For example, we considered KS2_READSCORE to be missing if KS2_READOUTCOME = 'B', i.e. the pupil was working below the level of the test.

Baseline measures and covariates

The NGRT was taken by all pupils in the trial at baseline before treatment status was allocated to schools. We used the baseline measures of the overall test score, sentence completion, and passage comprehension subscales to control for baseline attainment in their respective analyses.

For FSM eligibility, gender, EAL status, and Key Stage 2 SATs attainment, we used data from the NPD (as opposed to data from schools) where possible to minimise missingness. The following covariates were collected from the NPD:

- FSM status, obtained through the EVERFSM_6_P from the Spring Census 2023.
- EAL status, obtained through LanguageGroupMajor_SPR23 from the Spring Census 2023.
- Pupil gender through KS2_SEX.

We obtained the following variables for the pupil characteristics section of the IPE:

- SEND status (used for the IPE) used KS2_SENF and KS2_SENTYPE from Key Stage 2 dataset.
- Ethnicity through EthnicGroupMinor_SPR23 and EthnicGroupMajor_SPR23 from the Spring Census 2023.

If a variable was missing in the NPD but not in the school records, we used the value provided by the schools. This was done for 20 pupils who could not be matched to the NPD. For this group, we used the FSM status, gender, and EAL status reported

[9] 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.

by the schools. The data received from the schools was of high quality and there were no missing observations of FSM status.

Additionally, gender data was not available in the NPD dataset for all Year 5 pupils, so we used the pupil gender reported by the schools.

In the case of EAL status, a limited number of records in the NPD did not match the value in school data. In those cases, the NPD value took preference.[10] In the case of gender, we assume the gender reported by school was closer to the pupil's gender expression than the one recorded in KS2_SEX.

Sample size

The sample size was determined by the effect sizes achieved in the efficacy trial. To detect the effect on the primary outcome that was found in the efficacy trial (0.14 in Hedges' g), we estimated that we would need a sample of 257 schools. For the subgroup analysis, we would need 179 schools to detect the effect found in the efficacy trial on FSM pupils (which was 0.20 in Hedges' g).

Together with the delivery team, we decided to be conservative and aim for a higher sample of 300 schools with two goals: i) to allow for the possibility of not hitting the school recruitment target; and ii) to account for smaller effect sizes than those found in the efficacy trial (as we would expect from a scaled-up intervention).

Power calculations were conducted using R.

A first batch of 225 schools was randomised in July 2023 and a second batch of 70 schools in September 2023. The total number of successfully recruited schools for the trial at that point was 295. In total, 148 schools were allocated to the treatment group and 147 to the control group, with 2,118 pupils in the treatment group and 2,145 in the control group, making a total of 4,263 pupils.[11] 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.

Based on this sample, we concluded that the trial therefore, was well powered to detect an effect on the whole sample and the FSM subgroup.

Two of the estimated parameters at the protocol and randomisation stages were substantially different in size when calculated from the trial data. The pre-/post-test correlations were about 1.6 times smaller than estimated, and the ICC for the FSM subgroup was also about 1.6 smaller than estimated. Overall, this had a very limited effect on the power calculations. The minimum detectable effect size (MDES) estimated in the protocol were very similar to those calculated from the trial data.

[10] We have assumed that the NPD contains a more accurate reflection of the official definition of EAL, as schools might more loosely define it in their management information systems (MISs) and/or have less robust QA of their in-house data. However, it is also possible that the school data is more accurate. For example, a discrepancy between school and NPD data could come from the school's MIS data being more up to date—a pupil's EAL status could be identified after census day, be updated almost immediately in the school's MIS, but not updated in the NPD until the following census. Either way the numbers are so small that this choice is unlikely to make any substantive difference to the results.

[11] 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.

Table 7: MDES at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.13	0.16	0.13	0.16	0.12	0.15
Pre-/post-test correlations	Level 1 (pupil)	0.56	0.58	0.56	0.58	0.36	0.35
	Level 2 (class)	–	–	–	–	N/A	N/A
	Level 3 (school)	–	–	–	–	N/A	N/A
Intracluster correlation coefficients (ICCs)	Level 2 (class)	–	–	–	–	N/A	N/A
	Level 3 (school)	0.15	0.18	0.15	0.18	0.15	0.11
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		12	6	14	6	13	5
No. of schools	Intervention	150	150	148	148	145	140[a]
	Control	150	150	147	147	147	143[a]
	Total:	300	300	295	295	292	283
No. of pupils	Intervention	1,800	900	2,118	829	1,971	748
	Control	1,800	900	2,145	848	1,907	777
	Total:	3,600	1,800	4,263	1,677	3,878	1,525

[a] *The difference in the number of schools between the full sample and FSM subsample is due to some schools not having any FSM-eligible pupil in their group, so the number of school clusters in the FSM subgroup analysis is smaller. N/A = not applicable.*

Randomisation

FFT recruited participants as described in the section on participant selection. Randomisation was done after baseline testing by BIT. It was done in batches to enable FFT to begin setting up delivery with those schools that were recruited and baselined before the summer holidays. FFT communicated the results of randomisation to schools. In the first batch, the researcher implemented simple randomisation to allocate 225 schools to treatment and control groups. A fixed random seed (130823) was set to ensure the allocation could be reproduced. Each school was assigned a random number between 0 and 1, which was then ranked from 1 to 225. The ranked list was split in half, with schools ranked 1–113 allocated to treatment and schools ranked 114–225 allocated to control. This resulted in 113 treatment schools and 112 control schools. Within each randomisation batch, no blocking or stratification was used—each school had an equal 50% probability of assignment regardless of school characteristics. The sample size of 225 schools was considered large enough to minimise the risk of chance imbalances in baseline characteristics between groups through simple randomisation. The process followed for the second batch of 70 schools was identical. See the randomisation code in Appendix F.

Researchers carrying out baseline testing were blind to allocation. Analysis was not undertaken blinded to randomisation but followed the pre-specified plan and was quality assured by researchers from outside of the project team.

Statistical analysis

All analyses were conducted with complete cases only (observations missing one or more values were dropped from the analysis). Analyses were conducted in Rstudio, using two-sided significance tests, at the 5% significance level, on an intention-to-treat (ITT) basis.

Primary analysis

Analysis of the primary outcome, the NGRT overall score, was carried out using an ordinary least squares (OLS) regression with clustered standard errors at the school level,[12] 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 (see the note below) post-test assessment month for individual i, in school s; and
- ε_{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 (GL Assessment, 2025b)[13] age, gender, and EAL status are consistently correlated with NGRT scores, so they were included in the model as individual-level covariates. This increased the precision of the treatment estimate by reducing the idiosyncratic variance and made the model more robust in case of spurious and moderate imbalances in these covariates after randomisation. Including these covariates in the primary model is contrary to the EEF guidance but was agreed as appropriate with the EEF on this occasion. To assess the robustness of this model, two other models were estimated, one without these covariates and another with treatment status as the only covariate, and the results of the main specification were compared (see the 'Additional analyses and robustness checks' section below).

[12] This is a different approach to the EEF's standard guidance, which is to use a multilevel model. We favoured OLS over a multilevel model in this case for two main reasons: first, the interpretation of the OLS model is more straightforward, and in this case we are only interested in the population average treatment effect; second, the added complexity of multilevel models allows for estimation of the within-cluster and between-cluster effects at the expense of having to make stricter assumptions about the exogeneity of random effects, while a clustered OLS is more robust to model misspecification. Given that in this trial we have a large number of clusters (295) of similar size, and we do not expect a lot of variance in the cluster size by design (as it is a targeted intervention and all schools had to keep their selected pupils between 12 and 16 in order to participate), the ability of both models to reject the null will be similar. Hence, we opted for OLS, and we obtain the ICC at the school level separately.

[13] According to the test developers (GL Assessment), female pupils perform better in the NGRT test than male pupils by an average of 3.1 Standard Age Score (SAS) points, and non-EAL pupils perform better than EAL pupils by an average of 3.6 SAS points. These differences are significant at all age groups.

In the previous efficacy trial of Reciprocal Reading, endline NGRT assessments were collected between May and July. To accommodate data collection from the increased sample of 292 schools in the current trial, however, the testing window was extended between 18 March 2024 and the end of the school year in July 2024. This approach aimed to balance capacity constraints and limit attrition, while keeping the time window as narrow as possible. This was to minimise the chance of differences in assessment timings between treatment and control schools, given that an additional month of schooling could already make a difference in academic attainment for children at this age. Starting earlier than 18 March 2024 could have meant that pupils had fewer Reciprocal Reading sessions before the endline assessment than the participants in the efficacy trial, and so the average treatment effect might have been smaller. To minimise this risk, assessments were conducted at least three weeks after a school had the second training session, in a staggered process, following the same order in which schools had started implementation.

As outlined in the Statistical Analysis Plan (Torres Blas and Taylor, 2019), since there were more than two months of difference between the first and the last endline assessment (the first was carried out in March and the last one in July), assessment month fixed effects were included as an additional covariate in the regression to capture the variance in outcomes due to having more or fewer months of schooling when the assessment was conducted. The assessment month was also included in the covariates that were assessed for imbalance at baseline, which showed a slight imbalance where the control group was assessed slightly earlier than the treatment schools.

Secondary analysis

NGRT passage comprehension subscale

Analysis was carried out using an 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;
- 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);
- PrePC_{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 post-test assessment month for individual i, in school s; and
- ε_{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 β_1 in this model could be biased because of sample selection. Pupils who do not score the minimum required points in the sentence completion section of the test are not given passage comprehension tasks, so missing data on NGRT passage comprehension may not be random. Since the treatment is expected to improve sentence completion abilities, pupils in the treatment group are more likely to achieve this improvement at the endline, and hence, more likely to have a passage comprehension score. A more detailed explanation of this issue can be found in the Trial Protocol (Cappellini *et al.,* 2022, pp. 18–19).

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

- reporting the percentage of pupils with an NGRT overall reading score that are missing the passage comprehension score, at baseline and endline;[14]
- reporting the number of pupils in the treatment and control groups that have no passage comprehension score at baseline but do at endline;
- 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 was 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 passage comprehension is missing at endline and 0 if not missing);
- p_{is} is the probability that a given observation is missing the NGRT passage comprehension 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.

Less than 5% of pupils with NGRT overall score at endline were missing a passage comprehension score, and the coefficient of treatment assignment in the logit model was not statistically significant at 5%, so we did not carry out any further analysis as the risk of bias was considered to be minimal.

NGRT sentence completion subscale

Analysis of the NGRT sentence completion score was using the same model but changing the baseline attainment measure to the baseline NGRT sentence completion score.

Key Stage 2 SAT reading score for Year 6 pupils

We also estimated the impact of the intervention on the Key Stage 2 SAT reading scores of the subsample of pupils that are in Year 6 at the start of the programme and did their Key Stage 2 SATs in May 2024.

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

$$KS2\_READSCORE_{is} = \beta_0 + \beta_1 T_{is} + \beta_2 PreNGRT_{is} + \beta_3 Batch_s + \epsilon_{is}$$

where:

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

- $KS2\_READSCORE_{is}$ is Key Stage 2 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);
- Pre$NGRT_{is}$ 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
- ε_{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).

We are limited in our ability to draw causal conclusions from this subgroup analysis, as the cluster size is around 50% smaller compared to the primary outcome analysis. However, it can provide indicative evidence of the impact of Reciprocal Reading on standardised reading attainment measures.

Analysis in the presence of non-compliance

Compliance was defined by a binary variable equal to 1 if a pupil attended at least 20 Reciprocal Reading sessions, and 0 otherwise, based on FFT's recommendation.[15] This was based on pupil attendance data that was collected routinely by the teachers leading the sessions and shared with FFT. The completion of pupil registers by teachers was routinely monitored by FFT, who shared the data with BIT at the end of the intervention.

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

Stage 1 (Equation 4)

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

Stage 2 (Equation 5)

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

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 first stage of the 2SLS of Equation 4;
- $X_{is}$is a vector of pupil and school-level covariates including year group, gender, EAL status, randomisation batch, and post-test assessment time for individual i, in school s; and

[15] 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 sessions includes a margin for pupils to miss up to four sessions.

- ε_{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 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 estimate, we ran sensitivity checks by re-estimating the model using different thresholds that fall just below 20 (18, 15).

We also conducted a second sensitivity analysis where we considered a pupil as compliant if they did the 20 sessions in a maximum of 12 weeks (excluding school holidays). This excludes 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.

Optimal dosage analysis (exploratory)

As a pre-specified exploratory analysis, we also investigated the optimal number of instructional sessions required to maximise gains in the NGRT score using the following analysis.

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 was not randomly assigned, a descriptive analysis was employed instead. The analysis examined the relationship between the number of sessions and their impact on reading comprehension outcomes.

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

1. There could be low variability in the number of Reciprocal Reading 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 three weeks after they have done their second training session. 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 the dosage.
2. Pupil attainment could affect attendance (e.g. lower attainers could choose 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 regressed the number of sessions on pupil individual characteristics (including baseline attainment) using the subsample of treatment pupils using OLS.[16] If a regression coefficient was statistically significant at the 5% level, that individual characteristic was considered a predictor of session attendance and evidence of endogeneity. These findings are used to moderate the causal interpretation of this analysis.

We then regressed the change between baseline and endline NGRT score (calculated as the raw difference in scores) on a set of binary indicators for the number of sessions taken and on individual characteristics (gender, FSM status, EAL status, and baseline NGRT score). We used five binary indicators that divided the distribution of sessions into five quintiles (1–20, 21–24, 25–27, 28–31, and 32–51). The indicators were equal to 1 if a pupil took a number of sessions that was in the corresponding quintile, and 0 otherwise, with the reference category being no sessions attended.

(Equation 6)

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

[16] The distribution of the number of sessions was zero-inflated, so we repeated the analysis using a quasi-Poisson model with very similar results.

where:

- W_i is a vector of quintile indicators w_1, w_2, 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 covariates: FSM status, gender, and EAL status; and
- ε_i is the heteroskedasticity-robust error term, for individual i.

The estimated coefficients for the W binary indicators suggest how much change in NGRT score is associated with being in that quintile of session attendance compared to attending no sessions, controlling for individual characteristics. The results are plotted in a graph, including the coefficient estimates and their confidence intervals (CIs). The percentile with the highest coefficient estimate has been considered as the optimal dosage, in the absence of evidence that pupils from a lower percentage achieved a similar change in NGRT scores. This optimal dosage analysis presents descriptive evidence that has been used to contextualise the findings for the CACE analysis and provide recommendations for FFT and schools on dosage.

Missing data analysis

Descriptive analysis

We produce 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.

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

Understanding patterns of missingness

As outlined in the Statistical Analysis Plan (Torres Blas and Taylor, 2019), given that the primary outcome is missing for more than 5% of the randomisation sample, we first try to identify the pattern of missingness, i.e. whether they are missing conditional on other covariates or outcomes, or not. This analysis is not done for missing covariates, as no covariates were missing for more than 5% of the sample with primary outcome data. We do not do any further robustness checks to account for missing covariates, given the low rates of missingness, apart from comparing the main results to a main model with no covariates.

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 depends on the context of the trial. Whenever possible, we 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 depended on the type of missingness and whether the missing data are covariates or outcome variables.

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

(Equation 7)

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

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 primary outcome (the endline NGRT overall test score);[17]
- 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.

P-values below 0.05 are considered evidence of missingness being conditional on covariates or MAR.

Robustness checks

We find the outcome is MAR conditional on identified and available covariates that are not in the main specification (FSM status), so we estimate a model including that covariates and interpret the results. We also estimate an unadjusted model and compare the results with the other models.

We also find evidence of differential attrition, so we use bounds analysis (Lee bounds) to determine an 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 depended on our confidence that the monotonicity assumption holds (Lee, 2009). We discuss this assumption in the results section.

Subgroup analyses

We conduct 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 8 below).
2. Estimate a similar model including an interaction term for FSM eligibility, using a pooled sample (see Equation 9 below).

[17] The Statistical Analysis Plan specification wrongly identified this as the 'KS2 maths score' (Torres Blas and Taylor, 2019). This was a typo that has been corrected here.

Both approaches 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 are compared and reported.

FSM eligibility has been derived from the variable "EVERFSM_6_P" in the NPD.

Model 1: Split sample estimation

(Equation 8)

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

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 vector of pupil covariates[18] including, year group, gender, EAL status, and if applicable, post-test assessment time for individual i, in school s; and
- ε_{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 has been estimated for the subsample of pupils with $EverFSM_{is}$ equal to 1.

The reported effect size for FSM-eligible pupils is the treatment coefficient in Equation 8 (δ_1), in Hedges' g.

Model 2: Interaction effect

(Equation 9)

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

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 six years and 0 if not;

[18] The Statistical Analysis Plan specification wrongly included 'FSM status' in this list (Torres Blas and Taylor, 2019). This was a typo that has been corrected here.

- $PreNGRT_{is}$ 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
- ε_{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 β_3, its SD and p-value are reported, emphasising the range of effects that are compatible with the 95% CI.

As a sensitivity check, we also compute the *ES* for FSM-eligible pupils from Model 2 and compare it with Model 1. The effect size from Model 2 has been calculated as the sum of β_1 (the coefficient from the ITT variable) and the interaction effect β_3. This sum should be analogous to the treatment coefficient found in the split sample Model 1, δ_1.

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

We explore any differences and discuss their implications for the results.

Additional analyses and robustness checks

Models without covariates

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

Additionally, the primary outcome is also re-estimated with a model without pupil-level covariates, only the treatment assignment, baseline attainment, and trial design characteristics (the randomisation batch). This model allows for more comparability with other trials by the EEF and will be used to see if the inclusion of individual-level covariates in the main specification changes the results meaningfully.

(Equation 10)

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

where:

- Y_{is} is the endline NGRT overall score for individual i, in school s;
- 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$NGRT_{is}$ 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
- ε_{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).

Not pre-specified exploratory analysis

Sensitivity analysis: IT issues

We re-estimate the primary outcome model excluding all pupils from both treatment arms that were flagged by Qa Research as having experienced significant IT issues during endline testing. We compare the results with the main model to assess the extent to which these issues could have affected the main results.

Sensitivity analysis: Excluding month of test

We re-estimate the primary outcome model excluding the month of endline testing as a covariate and compare the results with the main model to assess the risk of bias from the differences in the month of test.

Estimation of effect sizes

The effect size for the primary and secondary analysis, including the FSM subgroup analyses, is 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^2 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^2 is the same for the control group.

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

We report 95% CIs for the effect sizes of all primary and secondary outcome analyses.

The passage comprehension and sentence completion subscales are 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 significance thresholds for the p-values reported for the two secondary analyses on NGRT sentence completion and NGRT passage comprehension scales are adjusted for multiple comparisons using the Benjamini-Hochberg method.

By contrast, we do not adjust the p-value of the other secondary outcome (Key Stage 2 SAT reading results). The analysis on Key Stage 2 SAT reading test results is done with a sample subgroup so it is not an equivalent comparison. The interpretation of the results for this outcome is distinct from the two NGRT subscales.

Estimation of ICC

ICCs for the primary outcome (NGRT overall score) are calculated at the school level at pre-test and post-test. We also calculate the ICC for the Key Stage 2 SAT reading scores at post-test, as Key Stage 2 SAT scores were only available for Year 6 pupils.

We estimate a one-way random effects analysis of variance (ANOVA) model with the school as a random effect (see Equation 11 below). The ICC will be calculated from the different variance components derived from the model.

(Equation 11)

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

where:

- Y_{is} is the outcome for individual i, in school s;
- μ 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_a^2 (the outcome variance at the school level) and the variance of e_{is} is denoted s_e^2 (the residual variance). The school ICC will be defined as:

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

This 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.

IPE

The IPE used a mixed methods approach, combining monitoring and administrative data with quantitative data from surveys and qualitative information from observations, interviews, and focus groups. A key component of the qualitative work was in-depth case studies of four schools that were implementing the intervention. The case studies combined semi-structured observations of Reciprocal Reading sessions with interviews with teachers and focus groups with pupils. In addition to the case studies, qualitative data was collected from parents of pupils receiving the intervention. The quantitative and qualitative data were combined following a convergent parallel approach (Fetters *et al.,* 2013) whereby data was collected in a similar time frame, analysed separately, and integrated during the interpretation of the findings phase. This design allowed the researchers to compare, contrast, and/or combine the different sources, triangulating them to understand the implementation processes more comprehensively. The details of each method are explored in turn below.

Methods

Monitoring data

FFT's data team shared data on the number of sessions delivered in schools and pupil attendance. This was collected through FFT's digital platform, where teachers/teaching assistants and school coordinators completed session registers on a regular basis (they were encouraged to log session attendance weekly). The data was shared with BIT at the end of programme delivery. However, the delivery team also monitored data collection periodically through check-ins to ensure quality and completeness of the data. As a result, BIT received dosage data from 100% of the schools in the treatment group that went through delivery (n=148), which included records for 1,930 pupils (all pupils who did the endline test, 92% of pupils

in the treatment group overall). Descriptive statistics on dosage are reported on these data in the 'Dosage' section of the findings. These data were also used in the compliance analysis of the impact evaluation.

Socio-demographic analysis

BIT collected pupil-level characteristics through the DfE's NPD including FSM status, EAL status, SEND status, SEND category, ethnicity, and gender. BIT accessed the NPD after the post-NGRT tests had been administered, using UPNs to match the pupils in the trial with the database. These same data fields were also shared with us by FFT, who collected them directly from schools. This helped us mitigate any risks of missing data, incomplete fields on the NPD, or other unexpected circumstances, which could have impacted the analysis. These data cover 4,234 pupils (99%).

FSM status was used for the subgroup analysis in the impact evaluation. The other data were used to describe the sample's composition and compare it to the general population of English school pupils and the sample from the efficacy trial. This analysis helps us to understand the generalisability of the impact findings. The results are reported in the 'Reach' subsection in the 'IPE results' section below.

School data analysis

Using the DfE's 'Get Information about Schools' portal and the DfE's 'School Census', we collected quantitative data on schools in the treatment and control groups, including % of FSM, location, and the Office for Standards in Education, Children's Services and Skills (Ofsted) rating. BIT's evaluation team obtained these data for all 295 schools in the trial. These data were used to compute descriptive statistics on the trial sample and compare it to the general population of English schools, and the sample from the efficacy trial. This analysis helps us to understand generalisability and could provide potential explanations for any differences in effects between the effectiveness and efficacy trials.

Observation of training sessions

We observed the delivery of six training sessions for school coordinators, teachers, and teaching assistants, delivered by FFT. Four sessions were observed between October 2023 and November 2023, and another two sessions from January 2024 to February 2024. All observations were in person and each at a different location and with a different trainer. A semi-structured observational framework was co-developed with FFT to ensure that all relevant dimensions of the training were captured. BIT researchers used this framework to check that the sessions met fundamental expectations of quality (as opposed to being a comprehensive assessment of quality). This was an area of particular interest as new trainers were trained up for the effectiveness trial. In previous trials, the training was delivered by a small group of highly experienced trainers who designed the intervention.

Training survey

A pre- and post-survey completed by training participants (school coordinators and teachers/teaching assistants) at both training days was used to capture short-term training outcomes by measuring:

- participants' attitudes to the training;
- changes in self-reported confidence in using Reciprocal Reading techniques;
- changes in self-reported knowledge of Reciprocal Reading techniques;
- changes in knowledge of Reciprocal Reading (using a knowledge test as a direct measure);
- characteristics of the trainee, such as years of experience in teaching, highest level of teaching qualification, and attitudes towards reading;
- attendance at the training session (partial one day of training, or full two days of training); and
- attendance at the online support session.

The questionnaire was developed by BIT and shared with the EEF and FFT for feedback before implementation. We piloted it with one ex-primary teacher in BIT's education team to ensure it was clear and to estimate completion time. The survey was administered to all school coordinators and teachers/teaching assistants participating in both training days. The Day 1 survey achieved a sample of 376 responses at pre-training (100%) and 357 responses at post-training (95%). The Day 2 survey received 305 responses at pre-training (83%) and 311 at post-training (85%). The short online questionnaires were hosted on SmartSurvey and completed at the start and end of the training sessions to ensure high completion rates and reduce the impact of confounding variables. To capture medium-term learning outcomes, some of the survey measures were repeated in a retrospective survey at the end of the delivery period (see below).

Retrospective survey of school coordinators and teachers/teaching assistants

A survey of school coordinators and teachers/teaching assistants involved in delivering the intervention was used to gain insight into responsiveness, adaptations, moderators, and unexpected impacts on usual practice. Retrospective online surveys were shared with all school coordinators, teachers/teaching assistants in the treatment group schools. The surveys were administered at the end of the delivery period and included both close- and open-ended questions measuring the following:

- repeat measures from training survey to monitor medium-term learning outcomes;
- self-reported attitudes to Reciprocal Reading;
- perceived impact on pupils of Reciprocal Reading (with specific questions on the impact of the intervention on FSM pupils);
- barriers and enablers to quality implementation;
- how the intervention was received by pupils eligible for FSM;
- adaptations to the delivery of the intervention;
- perceived impacts of the intervention on usual practice, i.e. has Reciprocal Reading replaced any curricular activities or alternative interventions;
- any other reading interventions or activities delivered to the treatment group during the trial period (using the same questions as those in the control group survey for comparability—see below); and
- how Reciprocal Reading compares to other additional initiatives to promote literacy, such as free reading time.

The survey for treatment schools received 297 responses from 135 of the 148 schools (91% response rate). Around 121 schools had at least two individual responses from staff involved in the programme. To encourage completion, we used behaviourally informed communications[19] and offered a monetary incentive (prize draw with a £200 prize). This questionnaire was administered via email using SmartSurvey and designed to be as short and user-friendly as possible (no longer than 15 minutes of completion time). The survey was launched in April 2024, when the majority of the schools had already completed >12 weeks of intervention.

The questionnaire was designed in collaboration with trial partners and piloted before launch. We piloted it with an ex-primary teacher in BIT's education team to ensure it was clear and to estimate completion time. BIT also considered the findings of the qualitative case studies to inform the questionnaire's design.

[19] We drew on behavioural science concepts to write effective emails to promote survey completion, e.g. using social norms by highlighting the large portion of respondents that had already answered.

Retrospective survey of school staff in control schools

The retrospective survey with school staff in control schools monitored compliance by exploring any deviations from business as usual resulting from the trial. Teachers selected pupils for the trial by identifying children who were good decoders but poor comprehenders. There was a risk that, due to the selection of pupils and exposure to the concept of Reciprocal Reading through communications about the trial, some teachers in the control group would provide their pupils with extra support or even self-implement aspects of the intervention. The survey also captured any alternative literacy interventions that control schools implemented, which might have resulted in improvements in reading among control pupils. The survey included:

- Teacher/teaching assistant reports on any additional support that was provided to pupils selected to take part in the trial.
- Teacher/teaching assistant reports on current reading interventions delivered in schools (availability and perceived effect on pupil literacy).
- Teacher/teaching assistant reports on current reading activities as part of the school activities, e.g. free reading times or library visits (availability and perceived effect on pupil literacy).

The survey was a short online questionnaire focused on measuring deviations from business as usual as a result of the trial and capturing any concurrent interventions in the schools to help us interpret the impact evaluation results. The survey received 173 responses from 137 of the 147 control schools (93% response rate). In addition, two treatment schools that withdrew from the study (out of five) answered the control survey at our request, to maximise the information that we had on these schools. To encourage high response rates, we used behaviourally informed communication and offered a prize draw incentive worth £200.

This survey data was used to contextualise the findings from the impact analysis. To add this context, we conducted the following analysis on the data:

- reported descriptive statistics to show the number and percentage of participants that participated in additional interventions during the intervention period (broken down by activity type and assignment); and
- conducted balance checks (using a normalised differences approach) to see whether there were substantial differences on this point between the treatment and control groups.

We have not re-estimated treatment effects based on the outcome of this analysis because conditioning on post-treatment variables can introduce bias (Montgomery *et al.,* 2018).

Case studies

The case studies gathered in-depth qualitative data from four schools, combining school staff interviews, pupil focus groups, and semi-structured observations[20] of Reciprocal Reading sessions. We purposely sampled the schools to include three schools with greater-than-average % of FSM pupils, and one school with average % of FSM pupils.[21] The achieved sample is shown below in Table 8 alongside a breakdown of activities carried out in each school. In total, we carried out seven interviews with school staff (four with school coordinators, three with teachers/teaching assistants), four focus groups with pupils, and four observations during March 2024 – April 2024.

[20] The Trial Protocol describes these as 'structured observations' (Cappellini *et al.,* 2022). This is a typo as the observations were semi-structured.
[21] The average proportion of pupils eligible for FSM was 25.7% for the 2024/2025 academic year.

Table 8: Case study sample and activities

Case study	% FSM pupils	Activities
School 1	70%	1 x interview with school coordinator[a] 1 x pupil focus group (Year 6) 1 x observation (Year 6)
School 2	64%	1 x interview with school coordinator 1 x paired interview with session leads 1 x pupil focus group (mix of Year 5 and Year 6) 1 x observation (Year 5)
School 3	62%	1 x interview with school coordinator 1 x interview with session lead 1 x pupil focus group (Year 6) 1 x observation (Year 6)
School 4	23%	1 x interview with school coordinator 1 x paired interview with session leads 1 x pupil focus group (Year 5) 1 x observation (Year 5)

[a] *Only one school staff interview was carried out due to challenges recruiting the session lead to take part in an interview. However, the school coordinator in this school was also delivering the intervention so was able to comment on delivery.*

Data collection tools (observation proforma and topic guides for interviews and focus groups) were used to guide the interactions. We received input from FFT into the observation proforma to ensure we were capturing elements of fidelity and quality comprehensively. The focus groups and interviews were recorded and transcribed to aid analysis.

Parent interviews

We carried out remote (phone/video call) interviews with a sample of 20 parents of pupils in the treatment group. Interviews explored family-level attitudes to reading and pupils' home reading environment. Half of the sample (50%) were parents of children eligible for FSM to allow comparisons between the home reading environments of pupils eligible for FSM and non-eligible pupils and to shed light on how Reciprocal Reading may act as a gap closer. We also considered parents' gender as a monitoring criterion, recruiting at least 25% of the sample from the least represented gender. This meant that our sample consisted of 25% of fathers and 75% of mothers. These interviews were recorded if participants consented so that they could be revisited during the analysis phase. Participants were offered a £20 voucher as compensation for their time.

While findings from the interviews with parents helped to contextualise some of the findings from interviews with teaching staff about perceived outcomes, the data was more limited due to parents' low awareness of the programme and challenges recognising changes in their child's reading habits and abilities. They therefore feature less prominently in the findings than other methods.

Table 9: IPE methods overview

Research method	Data collection methods	Participants / data sources[a]	Data analysis	Implementation / logic model relevance	Research question
Monitoring data	Attendance and session register – collated onto FFT's portal	All 148 treatment schools that delivered the programme	Descriptive statistics, CACE analysis	Dosage, Compliance	2, 3, 4
Socio-demographic sample analysis	NPD	3,878 pupils (out of 4,263)	Descriptive statistics	Moderators	3
School data analysis	DfE's school data portal	All 295 schools	Descriptive statistics	Moderators, Contextual factors	3
Training observations	Semi-structured observations	Purposive sample of four sessions in September–October and another two in January–February	Observational framework	Moderators, Causal mechanisms, Adaptations	1

		(each by a different trainer)			
Training surveys (pre and post)	Online questionnaires	Day 1: 376/357 responses Day 2: 305/311 responses	Descriptive statistics	Moderators, Quality, Causal mechanisms	1
Retrospective survey of treatment schools	Online questionnaires	297 responses from 135 treatment schools	Descriptive statistics	Quality, Fidelity, Adaptations, Moderators, Usual practice	2, 5
Retrospective survey of control schools	Online questionnaires	173 responses from 137 control group schools	Descriptive statistics	Usual practice	7
Case studies	Semi-structured interviews, structured focus group discussion, semi-structured observations	Purposive sample of four schools (three with high % of FSM pupils and one with national average level)	Thematic analysis	Quality, Responsiveness, Fidelity	2, 3, 5, 6
Parent interviews	Semi-structured interviews	Purposive sample of 20 parents from the treatment group schools (ten whose children are eligible for FSM and ten whose children are not)	Thematic analysis	Moderators	5

[a] *Samples for the whole trial or entire arms do not consider attrition, which is expected to be up to ~20%.*

Analysis

Analysis of pupil-level characteristics

An analysis of pupil and school characteristics was conducted to understand the composition of the sample and how it compared to the wider population and the sample from the efficacy trial. We reported descriptive statistics for these data, including:

- Frequency, reported as a percentage.
- Minimum and maximum values, reported as a count.
- Averages, reported as mean and median.
- Distribution, reported as SD.

All the descriptive statistics for pupil and school characteristics are reported in tables, and bar charts to represent frequency.

Analysis of monitoring data

The analysis of monitoring data includes computing descriptive statistics to report on the intervention dosage delivered to pupils in the treatment group. The same descriptive statistics outlined for pupil and school characteristics were used (frequency, minimum and maximum, averages, and distribution).

Additionally, we examined differences in delivery at the regional level. We used the delivery clusters specified by FFT as the analytical unit and computed the frequency and mean number of sessions delivered in each cluster to monitor regional variations. These data are also used as part of the impact evaluation to estimate the CACE, which gives an indication of how dosage moderates the treatment effect.

Training survey analysis

To evaluate the training outcomes for school coordinators and teachers/teaching assistants, we conducted a pre- and post-analysis on the quantitative skills and knowledge items. The post-survey also included two to three open-text fields, which were coded using an inductive approach and semi-quantification. Using the SmartSurvey coding function, we coded each response with a theme and saw the % of responses that relate to a specific theme.

The following model was used to estimate the effects of the training on teachers' learning, using OLS regression. This pre- and post-analysis was conducted on an ITT basis, including all complete cases in the sample:

$$y_{it} = \alpha + \beta post_t + \gamma a_{it} + \gamma b_{it}{}' \varepsilon_{it}$$

Where:

- y_{it} is the outcome[22] for respondent i in period t;
- $post_t$ is a binary indicator of the sample (1 for endline, 0 for baseline survey);
- a_{it} is a binary variable recording whether the training was delivered by an experienced trainer or a newly trained trainer;[23] and
- b_{it} is a binary variable recording whether the respondent attended the training in full or partially.

Retrospective surveys analysis

To analyse the retrospective surveys administered to school staff in the treatment and control groups, we used a mixed methods approach. Quantitative results are presented using descriptive statistics, specifically the percentage of respondents who chose specific responses. Qualitative responses were coded in Excel using a mixture of inductive and deductive coding.

Qualitative data analysis

Qualitative data from the observations, interviews, and focus groups were analysed thematically using the Framework approach, which allows in-depth exploration of the data by case and by theme (Ritchie *et al.,* 2014). This consisted of creating a matrix in which to organise the data, based on the topic guides and observation proformas. Data was summarised and displayed in the matrix. This was followed by working through the managed data to draw out the range of behaviours, experiences, and views, while identifying similarities, differences, and links between them.

Costs

Information on the cost of the intervention was collected from a sample of 15 schools.[24] We used the 'ingredient method' as specified by the EEF guidelines (EEF, 2023) to create a proforma with input from the FFT delivery team. The proformas were then completed during remote semi-structured interviews with school coordinators.

For each type of cost, a mean average was calculated. The costs of years two and three were adjusted for inflation (excluding any costs that were considered to be a one-off). The total was then divided by three, to give an average cost per school per year. To determine the cost per pupil per year, the average cost per school per year was then divided by 36 (i.e. the number of pupils estimated to take part in the intervention over three years).

[22] See description of outcomes above.
[23] An experienced trainer is defined as a trainer who has delivered Reciprocal Reading training as part of a previous delivery and was not trained for the purpose of this trial.
[24] This was the maximum number feasible within the budget for the methods proposed.

Timeline

Table 10: Timeline

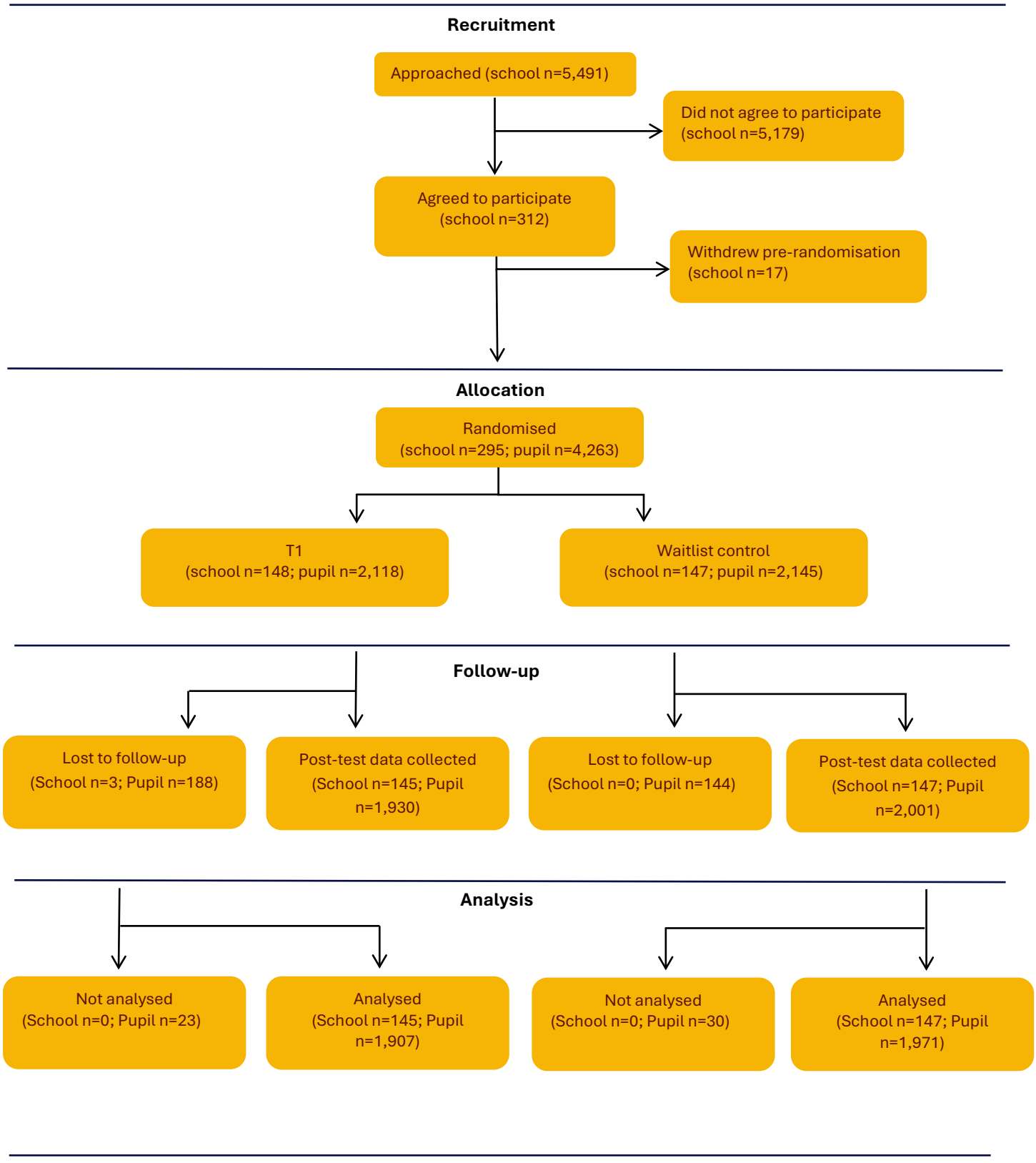
Dates	Activity	Staff responsible/ leading
14/11/2022	Ethical approval for the trial	Head of quantitative research at BIT
01/02/2023	Finalise recruitment documents	The EEF/BIT/FFT
06/02/2023	Start of school recruitment (FFT requests pupil-level data)	FFT
13/02/2023 – 17/02/2023	Spring half-term holidays	School dates
13/03/2023	Share Trial Protocol with trial partners	BIT evaluation manager / principal investigator
20/03/2023	Qa Research and testing process is introduced to schools during school leads webinars	FFT
03/04/2023	BIT starts receiving school data from FFT	FFT
03/04/2023 – 14/04/2023	Easter break	School dates
14/05/2023 – 15/10/2023	Evaluators enrol pupils on the GL Assessment platform	BIT evaluation manager
14/04/2023 – 01/09/2023	Evaluators book Qa Research school visits	Qa Research
09/05/2023 – 12/05/2023	Key Stage 2 SATs	School dates
22/05/2023 – 15/10/2023	FFT shares pupil data with the evaluators	FFT
29/05/2023 – 02/06/2023	Summer half-term holidays	School dates
12/06/2023	Start administering NGRT tests in schools	BIT / Qa Research
12/06/2023 – 21/07/2023	Administering 80% of NGRT tests in schools (Qa Research visits schools)	BIT / Qa Research
07/07/2023	Publish Trial Protocol	The EEF / BIT
21/07/2023	Trial registration on the OSF	BIT
21/07/2022	Randomisation of the first batch of schools	BIT
21/07/2022	FFT finishes recruiting schools	FFT
04/05/2023 – 21/07/2023	Online briefing with school coordinators	FFT
30/07/2023	Evaluators to share draft training surveys with trial partners	BIT
21/07/2023 – 04/09/2023	Summer holidays for schools	School dates
05/09/2023	FFT shares the final list of schools that will participate in the trial with the evaluators	FFT
04/09/2023 - 16/10/2023	Administer the remaining 20% of NGRT test results	BIT / Qa Research
15/09/2023	Randomisation (second batch)	BIT
01/10/2023 – 30/11/2023	Reciprocal Reading Day 1 training delivered to schools (and pre- and post-training surveys collected)	FFT
01/10/2023 – 11/03/2024	Reciprocal Reading delivery in schools	FFT / partner schools
23/10/2023 – 27/10/2023	Autumn half-term holidays	School dates
01/12/2023 – 16/03/2024	Recruiting case study schools	BIT
21/12/2023 – 08/01/2024	Christmas holidays in school	School dates
12/02/2024	Evaluators share qualitative research instruments with project partners (observational frameworks and interview guides for pupils, school staff, and parents)	BIT
12/02/2024 – 16/02/2024	Spring half-term holidays	School dates
16/02/2024	Share Reciprocal Reading delivery timetable with Qa Research	FFT
26/02/2023 – 25/03/2024	Qa Research to book school visits for the retrospective NGRT assessments	Qa Research

Dates	Activity	Staff responsible/ leading
09/03/2024	Statistical Analysis Plan submitted to the EEF	BIT quantitative researcher
04/03/2024 – 08/07/2024	Retrospective NGRT tests administered in schools	BIT / Qa Research
11/03/2024	Most schools completed ≥12 weeks of intervention	FFT / partner schools
21/03/2024 – 03/04/2023	Co-development of parent recruitment strategy for parent interviews	BIT / FFT
29/03/2024 – 12/04/2024	Easter holiday in schools	School dates
04/03/2024 – 28/06/2024	Conduct qualitative research in case study schools	BIT qualitative researcher
13/05/2024 – 19/05/2024	Key Stage 2 SATs	School dates
04/03/2024 – 08/07/2024	Administer post-NGRT tests in schools (Qa Research visits schools)	BIT evaluation manager / Qa Research
10/04/2024 – 06/05/2024	Conduct parent interviews (x20)	BIT qualitative researcher
15/05/2024	Evaluators share draft retrospective surveys for school staff (control and treatment arm)	BIT evaluation manager
31/05/2024	Evaluators launch retrospective survey with school staff (treatment and control group)	BIT quantitative researcher
27/05/2024 – 31/05/2024	Summer half-term holidays	School dates
05/04/2024 – 28/06/2024	Delivery team conducts second meeting with schools	FFT
01/06/2024	Submit NPD data access request	BIT quantitative researcher
03/06/2024 – 21/06/2024	Cost evaluation interviews with school coordinators	BIT qualitative researcher
24/07/2024 – 02/09/2024	Summer holidays	School dates
30/09/2024	NPD data release (unamended dataset)	DfE
01/01/2025	Analysis in ONS SRS begins	BIT quantitative researcher
25/07/2025	First draft of report submitted to the EEF	BIT evaluation manager /principal investigator

Impact evaluation results

Participant flow including losses and exclusions

Figure 4: Participant flow diagram (two arms)



Attrition

Table 11 summarises attrition for the primary outcome (the NGRT overall score). The number of pupils analysed refers to those with both outcome and complete covariate data. Overall attrition was less than 10%, which is good considering that all outcome data had to be collected directly from pupils through a non-standard assessment. Attrition of less than 10% is the EEF's highest standard for completeness of outcome data.

Table 11: Pupil-level attrition from the trial (primary outcome)

		Intervention	Control	Total
No. of pupils	Randomised	2,118	2,145	4,263
	Analysed	1,907	1,971	3,878
Pupil attrition (from randomisation to analysis)	Number	211	174	385
	Percentage	9.96%	8.11%	9.03%

Observations were missing from the primary analysis for the following reasons, listed in order of frequency. These reasons were reported to the researchers by the schools and by Qa Research (the company that delivered the baseline and endline tests in the schools). The number of pupils in each category and percentage over all attrition is provided in brackets (N = 385).

- Pupil left the school: 137 (35.6%).
- Reason not known: 96 (24.9%).
- Pupil missing at least one covariate: 53 (13.8%).
- Entire school withdrew from trial: 38 (9.87%).
- Pupil withdrawn from trial by parents: 29 (7.53%).
- Pupil removed from the programme and test by the teacher/the school: 20 (5.19%).
- Pupil persistently absent and consequently missed assessment: less than ten (percentage censored).
- Pupil did not assent to testing: less than ten (percentage censored).

Five schools in the treatment group withdrew from the trial at the start, citing lack of sufficient resources and staff as main reasons. While two of them agreed to do the NGRT tests (which could dilute the treatment effect), the outcome is missing for the pupils of the other three schools. A total of 20 pupils were removed from the programme by their school because the teachers decided that the programme was not appropriate for them; for example, because they could not follow the sessions sufficiently. These pupils were also removed from the endline assessment, contrary to the Trial Protocol (Cappellini *et al.,* 2022). This could have introduced some upward bias in the treatment effect as we expect similar pupils in the control group to have taken the endline assessment. We also find some evidence of differential attrition in the 'Missing data analysis' section below: assignment to treatment is a statistically significant determinant of missing NGRT outcome. Overall, the total attrition is low, there is not a large imbalance in attrition between trial groups and the covariates used in the analysis should absorb at least some of the imbalance, so this is not a large threat to internal validity.

Pupil and school characteristics

Table 12 summarises the baseline pupil-level characteristics of intervention and control pupils as randomised. In general, it shows a very high level of balance between groups, with negligible differences. It also shows that, compared to the national pupil population in the same age group, the trial sample had a substantially higher proportion of pupils eligible for FSM (40.1% vs 30.5%). It also had a higher proportion of EAL pupils (27.95% vs 23.4%).

Table 12: Baseline characteristics of pupils as randomised

Pupil level (categorical)	National-level mean[a]	Intervention group		Control group		Effect size (normalised difference in Hedges' g)
		n / N (missing)	Count (%)	n / N (missing)	Count (%)	
FSM eligibility = Yes	30.5%	2,106 / 2,118 (12)	842 (40%)	2,128 / 2,145 (17)	855 (40.2%)	-0.004
EAL status = Yes	23.4%	2,106 / 2,118 (12)	579 (27.5%)	2,128 / 2,145 (17)	605 (28.4%)	-0.021
Gender = female	49.1%	2,106 / 2,118 (12)	1,005 (47.7%)	2,128 / 2,145 (17)	1,037 (48.7%)	-0.020
Year group = 5 (vs Year group = 6)	N/A	2,118 / 2,118 (0)	1,051 (49.6%)	2,145 / 2,145 (0)	1,064 (49.6%)	< 2.2e-16
Pupil level (continuous)	**National-level mean**	**n / N (missing)**	**Mean (SD)**	**n / N (missing)**	**Mean (SD)**	**Effect size (normalised difference in Hedges' g)**
NGRT overall reading score	N/A	2,051 / 2,118 (67)	277 (45.4)	2,072 / 2,145 (73)	278 (44.3)	-0.004
NGRT sentence completion score	N/A	2,051 / 2,118 (67)	286 (47.3)	2,072 / 2,145 (73)	287 (45.5)	-0.021
NGRT passage comprehension score	N/A	2,022 / 2,118 (96)	275 (48.3)	2,049 / 2,145 (96)	274 (49.2)	0.016

[a] *Taking figures for the 2024/2025 academic year from https://explore-education-statistics.service.gov.uk/. Calculating averages across Years 5 and 6 where the data exists, otherwise figures are for primary schools across all age groups. N/A = not applicable.*

The NGRT scores do not have comparable national averages as it is not a nationally administered test. However, GL Assessment does produce alternative statistics to show how individual pupils compare to their peers at the same age. This analysis is based on the scores collected from a sample of 20,000 pupils that was representative of the UK population in 2014. This sample differs from ours in that it is ten years old, covers all UK regions (as opposed to just England) and includes a representative number of independent schools (as opposed to just state schools). However, the bell curve of scores produced from the sample is checked each year against new NGRT test data, and the majority of the sample is still from English state schools. So, for the purposes of an approximate comparison, it is good enough.

When we compared our trial sample with this representative sample, we find that the average overall reading score for our sample is slightly below average for children of the same age (a standardised score of 93.4 vs an average of 100). Another way of expressing this is that slightly more than 60% of the national population will have higher scores than the average pupil in our sample at baseline. When we look at the subsections of the test, we find that the overall reading score in our sample is driven by slightly higher scores in sentence completion and slightly lower scores in passage comprehension, but the sample remains below average for both.

These differences are to be expected given the pupil selection process for the intervention and are evidence that pupil selection was carried out as intended. This is discussed more in the IPE findings on reach.

Table 13 presents the analogous balance characteristics for the groups as analysed for the primary outcome. The results of these checks are almost identical to those for the randomised sample, showing a high level of balance across all baseline characteristics.

We have added one variable to the balance checks for the analysed sample: Month of test. This is the month in which the pupil took the NGRT endline test, taking the values of 1 to 12 to correspond with the calendar months. It is not a baseline covariate, but we have checked the balance to assess the risk of bias. The average month of testing was 5.12 in the intervention group, meaning that pupils in the group took the test, on average, just after the beginning of May. This figure in the control group was 4.95, meaning that pupils in the group took the test, on average, at the end of April. This mean difference of 0.17 months is equal to about five days. We do not expect this small difference in the average endline test date to introduce substantial bias but, as pre-specified, month of test was included as a covariate in the analysis to account for the spread of dates over which pupils took the test. All outcome analyses also controlled for year group, gender, and EAL status.

Table 13: Baseline characteristics of pupils as analysed

Pupil level (categorical)	National-level mean	Intervention group		Control group		Effect size (normalised difference in Hedges' g)
		n / N (missing)	Count (%)	n / N (missing)	Count (%)	
FSM eligibility = Yes	30.5%	1,907 / 2,118 (0)	748 (39.2%)	1,971 / 2,145 (0)	777 (39.4%)	-0.004
EAL status = Yes	23.4%	1,907 / 2,118 (0)	516 (27.1%)	1,971 / 2,145 (0)	554 (28.1%)	-0.024
Gender = female	49.1%	1,907 / 2,118 (0)	914 (47.9%)	1,971 / 2,145 (0)	958 (48.6%)	-0.014
Year group = 5 (vs Year group = 6)	N/A	1,907 / 2,118 (0)	937 (49.1%)	1,971 / 2,145 (0)	965 (49.0%)	0.004
Pupil level (continuous)	**National-level mean**	**n / N (missing)**	**Mean (SD)**	**n / N (missing)**	**Mean (SD)**	**Effect size (normalised difference in Hedges' g)**
NGRT overall reading score	N/A	1,907 / 2,118 (0)	278 (45.2)	1,971 / 2,145 (0)	278 (44.2)	0.005
NGRT sentence Completion score	N/A	1,907 / 2,118 (0)	287 (47.2)	1,971 / 2,145 (0)	287 (45.2)	-0.016
NGRT passage comprehension score	N/A	1,880 / 2,118 (0)	275 (48.0)	1,950 / 2,145 (0)	274 (49.3)	0.028
Month of test	N/A	1,907 / 2,118 (0)	5.12 (0.90)	1,971 / 2,145 (0)	4.95 (1.03)	0.174

N/A = not applicable.

Table 14 summarises the baseline school-level characteristics and how they compare to national-level figures. Around 90% of schools in the sample were urban, compared to 71% of schools at the national level. There was a substantially higher proportion of schools rated as Good by Ofsted in the trial sample (72% vs 66% in England). The sample schools also had a higher-than-average proportion of pupils eligible for FSM (32% vs 24%). Key Stage 2 SAT reading data suggests that trial schools had similar intakes to the national average when it comes to reading attainment.

Table 14: Baseline characteristics of schools as randomised

School level (categorical)	National-level mean	Intervention group (N = 148)		Control group (N = 147)	
		n	Count (%)	n	Count (%)
Ofsted rating					
Good	66.1%[a]	111	75%	102	69.4%
Outstanding	11%	16	10.8%	20	13.6%
Requires improvement	6.4%	11	7.4%	17	11.6%
Serious weaknesses	0.2%	2	1.4%	–	–
Special measures	0.2%	1	0.7%	–	–
N/A	15.9%	7	4.7%	8	5.4%
Type of school					
Academy converter	26.5%	34	23%	47	32%
Academy sponsor led	9.1%	20	13.5%	20	13.6%
Community school	27.4%	62	41.9%	48	32.7%
Foundation school	2.7%	4	2.7%	4	2.7%
Free school	1.6%	1	0.7%	3	2%
Voluntary aided school	11.3%	18	12.2%	14	9.5%
Voluntary controlled school	7.8%	9	6.1%	11	7.5%
Location					
Rural	26.8%	18	12.2%	13	8.8%
Urban	70.8%	130	87.8%	134	91.2%
N/A	2.4%	0	–	0	–
School level (continuous)	National-level mean	N (missing)	Mean (SD)	N (missing)	Mean (SD)
Key Stage 2 SAT reading scaled score	105 (2022/2023)[b]	148 (0)	104 (2.87)	147 (0)	104 (2.88)
% of Key Stage 2 pupils meeting the expected standard in reading	73% (2022/2023)[c]	148 (0)	72.1% (13.9)	147 (0)	71.5% (14)
% of FSM pupils	23.8% (2022/2023)[d]	148 (0)	32.4% (17.7)	147 (0)	32.0% (16.6)

[a] *The numbers of national-level means for Ofsted rating, type of school, and urban or rural area, are from the authors' own calculations using the 2022/2023 DfE school census. It includes only the 19,650 primary schools in England that were open in that academic year.*
[b] *Source: DfE official statistics for Key Stage 2 attainment in academic year 2022/2023. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2022-23.*
[c] *Source: DfE official statistics for Key Stage 2 attainment in academic year 2022/2023. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2022-23.*
[d] *Source: DfE official statistics, for schools, pupils, and their characteristics in academic year 2022/2023. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2022-23.*

In sum, compared to the national average, the analysis sample is more urban, rated better by Ofsted and had a more economically deprived intake of pupils.[25] The location, school performance, and intake could moderate the effects of the intervention. The effects identified in our analysis may therefore, be different if the treatment were scaled up to all schools in England. The IPE findings suggest that the quality of the trainers, the teachers, and teaching assistants who deliver the programme are important factors in the success of the programme. Recruiting high-quality staff is probably easier in urban areas where the population density is higher (though this is speculation and we did not collect data on the topic in the IPE). So, if the programme were scaled-up nationally, an increase in the proportion of rural schools in the sample may lead to a reduction in the effects.

Outcomes and analysis

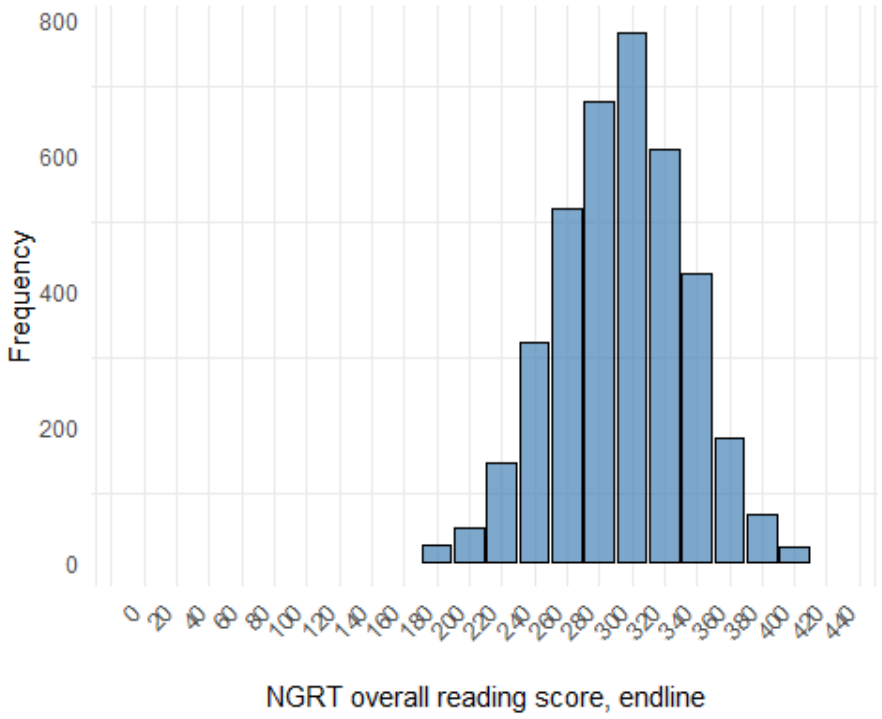
Primary analysis

The primary outcome measure was the NGRT overall reading score, a measure of overall reading proficiency. The NGRT is a standardised assessment that measures skills in sentence completion and reading comprehension and, unlike other national reading assessments like the Key Stage 2 SAT reading test (which is done in May of Year 6), it can be taken by Year 5 and Year 6 pupils immediately after the intervention. Scores can theoretically range between 0 and 500 points, so the estimated effect can take theoretical values between -500 and 500. Effects are presented as Hedges' g to make it easier to compare between outcomes and with other studies.

Figure 5 shows the distribution of outcomes. Scores are distributed across almost the full range (0 to 440). There is some right skew, but no evidence of floor or ceiling effects. The mean score at endpoint was 303.1, SD 45.53 (N = 3,878).

All counts from 0 to 180 and from 420 to 440 were non-zero but lower than ten and have been censored to protect the privacy of the pupils, following the ONS SRS guidance.

Figure 5: Histogram of the endline NRGT overall reading score



[25] This is partly explained by the fact that the trial was funded through the DfE's Accelerator Fund, which required that school recruitment was focused on the DfE's Education Investment Areas (see: www.gov.uk/government/publications/education-investment-areas/education-investment-areas).

Table 15 presents the results of the analysis for the NGRT overall reading score outcome. The unadjusted mean for the score is 304.552 (95% CI: 302.605, 306.498) in the intervention group and 301.687 (95% CI: 299.590, 303.783) in the control group. After adjusting for covariates in the analysis model, the effect size, in Hedges' g, is estimated to be 0.055 (a small positive effect, equivalent to one month of progress). While this is the best estimate of the intervention's impact, the 95% CI suggests the true effect could range from negligible (zero months of progress) to a slightly larger positive effect (two months of progress).

Table 15: Primary outcome analysis results

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Hedges' g (95% CI)	P-value
NGRT overall reading score	1,907 (211)	304.552 (302.605, 306.498)	1,971 (174)	301.687 (299.590, 303.783)	3,878	0.055 (-0.014, 0.124)	0.117

Secondary analysis

We also estimated the impact of the programme on three secondary outcomes: reading comprehension skills (measured by the NGRT passage comprehension score); sentence completion skills (measured by the NGRT sentence completion score); and reading attainment (measured by the Key Stage 2 SAT reading score).

Figure 6 shows the distribution of **passage comprehension score**. Scores can theoretically range between 0 and 500 points, so the estimated effect can take theoretical values between -500 and 500. Scores are distributed across almost the full range (40 to 480). There is some right skew, but no evidence of floor or ceiling effects. The mean score at endpoint was 299.41, SD 47.2 (N = 3,793).

All counts from 40 to 180 and from 440 to 480 were non-zero but lower than ten and have been censored to protect the privacy of the pupils, following the ONS SRS guidance.

Figure 6: Histogram of the endline NRGT passage comprehension score

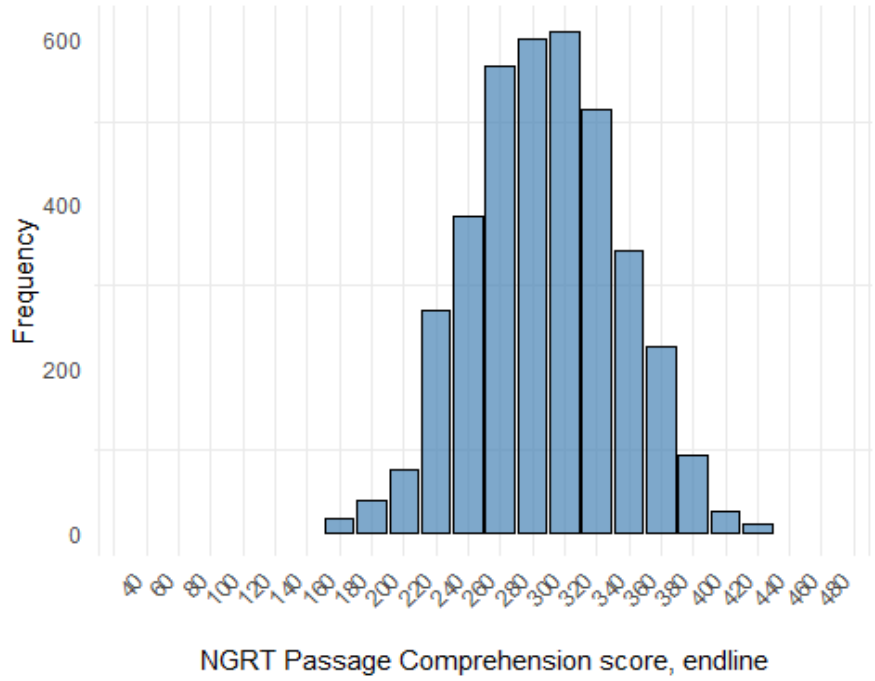


Figure 7 shows the distribution of **sentence completion score**. Scores can theoretically range between 0 and 500 points, so the estimated effect can take theoretical values between -500 and 500. Scores are distributed across almost the full range (0 to 480). There is some right skew, but no evidence of floor or ceiling effects. The mean score at endpoint was 314.66, SD 47.16 (N = 3,878).

All counts from 0 to 180 and from 440 to 480 were non-zero but lower than ten and have been censored to protect the privacy of the pupils, following the ONS SRS guidance.

Figure 7: Histogram of the endline NRGT sentence completion score

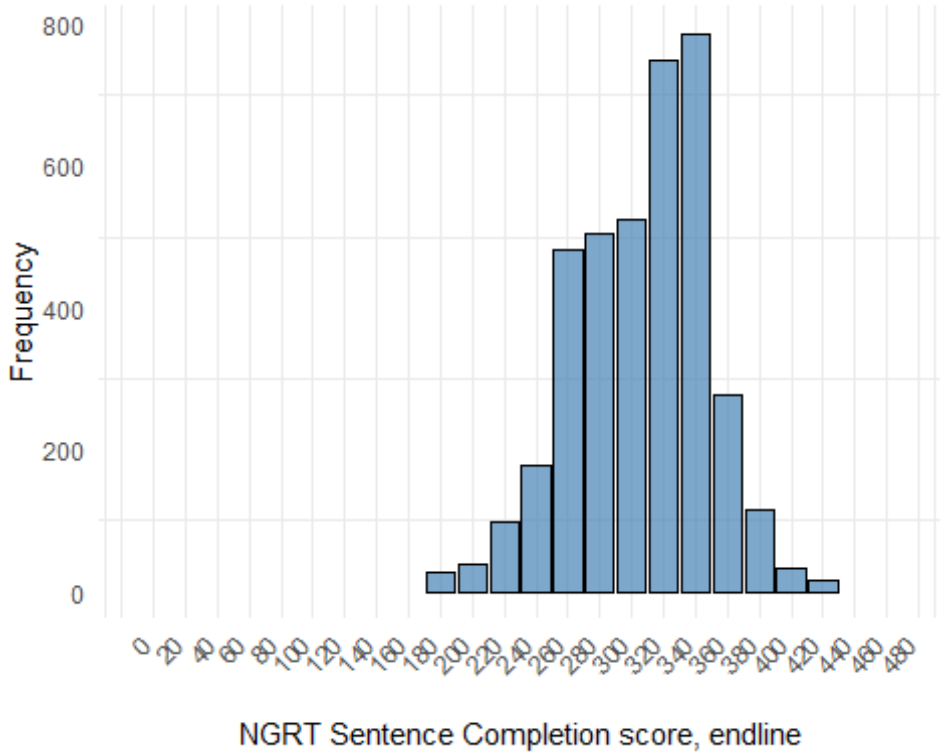


Figure 8 shows the distribution of **Key Stage 2 SAT reading score**. The scaled scores can theoretically range between 80 and 120, so the estimated effect can take theoretical values between -40 and 40. Scores are distributed across the full range (80 to 120). There is no strong evidence of skew, and no evidence of floor or ceiling effects. The mean score at endpoint was 101.18, SD 6.32 (N = 2,050).

All counts from 80 to 84 and from 116 to 120 were non-zero but lower than ten and have been censored to protect the privacy of the pupils, following the ONS SRS guidance.

Figure 8: Histogram of the endline Key Stage 2 SAT reading score

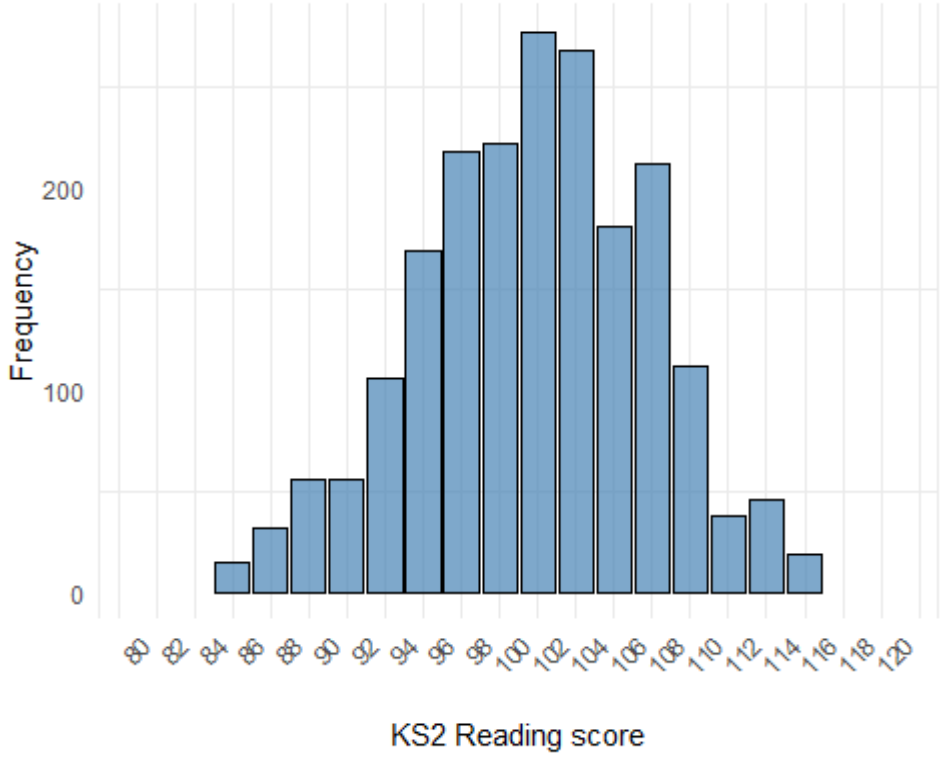


Table 16 presents the results of the analysis for the secondary outcomes.[26]

Table 16: Secondary analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Hedges' g (95% CI)	P-value
NGRT passage comprehension score	1,869 (249)	300.456 (298.369, 302.542)	1,924 (221)	298.397 (296.239, 300.556)	3,793	0.027 (-0.045, 0.099)	0.458
NGRT sentence completion score	1,907 (211)	315.927 (313.887, 317.967)	1,971 (174)	313.427 (311.275, 315.578)	3,878	0.052 (-0.017, 0.121)	0.139

[26] As per the Statistical Analysis Plan (Torres Blas and Taylor, 2019), the threshold by which we assess the significance of the two NGRT subscores should be adjusted to account for multiple comparisons, using the Benjamini-Hochberg procedure. The procedure works as follows: take the p-values from each comparison and arrange them in ascending order. Conventionally, we might compare these against a fixed significance threshold (usually 0.05) and anything smaller would be deemed significant. The Hochberg procedure instead compares them with a linearly increasing vector from 0.05/k (where k is the number of comparisons) to 0.05. Once a comparison is found not significant, all remaining comparisons are also classified as non-significant. In this case, the p-values for both subscores are already larger than 0.05 so the adjustments do not change the conclusion that estimated effects are not significant at conventional levels. However, in the reporting of results we avoid discussion of precision in terms that suggest that significance is a binary concept ('significant' or 'not significant') as this would be a mistake. Instead, we report the range of possible values with which the estimated effects are consistent.

Key Stage 2 SAT reading score	1,015 (1103)	100.995 (100.597, 101.393)	1,035 (1110)	101.365 (100.989, 101.741)	2,050	-0.049 (-0.168, 0.070)	0.424

NGRT passage comprehension score

For the passage comprehension score, the unadjusted mean is 300.456 (95% CI: 298.369, 302.542) in the intervention group and 298.397 (95% CI: 296.239, 300.556) in the control group (Table 16). After adjusting for covariates in the analysis model, the effect size, in Hedges' g, is estimated to be 0.027 (equivalent to zero month of progress) (Table 16). While this is the best estimate of the intervention's impact, the 95% CI suggests the true effect could range from a small negative (one month's less progress) to a small positive effect (two months of progress). This reasonably high level of uncertainty makes it hard to conclude what the effect of the intervention was on this outcome but, whatever the direction of the effect, it is likely to be small.

There is a theoretical risk of bias in the NGRT passage comprehension score, as pupils need to show a basic level of reading to access the passage comprehension part of the test (see page 11 of the Statistical Analysis Plan for more information; Torres Blas and Taylor, 2019). To assess the likelihood of such bias being introduced in the trial, we first calculated the percentage of pupils with an NGRT overall reading score that are missing the passage comprehension score, at baseline and endline:

- Endline: 1.1% (43 observations of 3,888).
- Baseline: 1.3% (52 observations of 4,071).

We then calculated the number of pupils in the treatment and control groups that have no passage comprehension score at baseline but do at endline:

- Treatment: 48 (2.3%).
- Control: 47 (2.2%).

Finally, we ran a logistic regression to test whether the probability of not having an endline score is correlated with treatment assignment. The coefficient on the treatment indicator was 0.208 (p=0.092).

With only 1.1% of pupils with an NGRT endline score missing a passage comprehension score, and no significant difference in the probability of having a score, we conclude that there is little to no risk of bias in the passage comprehension results.

NGRT sentence completion score

For the sentence completion score, the unadjusted mean is 315.927 (95% CI: 313.887, 317.967) in the intervention group and 313.427 (95% CI: 311.275, 315.578) in the control group (Table 16). After adjusting for covariates in the analysis model, the effect size, in Hedges' g, is estimated to be 0.052 (equivalent to one month of progress) (Table 16). While this is the best estimate of the intervention's impact, the 95% CI suggests the true effect could range from negligible (zero month of progress) to a small positive effect (two months of progress). The intervention seems therefore, to have had either no effect on this outcome, or a small positive one.

Key Stage 2 SAT reading score

For the Key Stage 2 SAT reading score, the unadjusted mean is 100.995 (95% CI: 100.597, 101.393) in the intervention group and 101.365 (95% CI: 100.989, 101.741) in the control group (Table 16). After adjusting for covariates in the analysis model, the effect size, in Hedges' g, is estimated to be -0.049 (equivalent to one month's less progress) (Table 16). While this is the best estimate of the intervention's impact, the 95% CI suggests the true effect could range from a small negative (two months less progress) to a small positive effect (one month of progress). This reasonably high level of uncertainty makes it

hard to conclude what the effect of the intervention was on this outcome but, whatever the direction of the effect, it is likely to be small.

Subgroup analyses

Subgroup analyses were conducted to establish whether treatment effects differed for pupils who are, and are not, eligible for FSM. Subgroup analysis was conducted both using an interaction effect and a model where the sample is limited to the subgroup of those who were eligible for FSM.

Table 17: Subgroup analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Hedges' g (95% CI)	P-value
NGRT overall reading score Only FSM-eligible pupils	748 (1,370)	298.241 (295.172, 301.309)	777 (1,368)	295.578 (292.182, 298.974)	1,525	0.035 (-0.059, 0.130)	0.465
NGRT overall reading score Interaction term	1,907 (211)	304.552 (302.605, 306.498)	1,971 (174)	301.687 (299.590, 303.783)	3,878	-0.038 (-0.142, 0.065)	0.468

As Table 17 indicates, when estimating the effect on the FSM subgroup only by restricting the sample ('Model 1'), those who were eligible for FSM saw effectively zero month of additional progress: treatment coefficient = 1.61, Hedges' g = 0.035 (95% CI: -0.059, 0.130).

In a second model ('Model 2') we included an interaction term. The coefficient on the interaction term (interpreted as the difference in effects between FSM and non-FSM pupils) is small (Hedges' g for the interaction effect model = -0.038; 95% CI: -0.142, 0.065)—equivalent to zero month of additional progress (Table 17). This is our best estimate of the difference in effects between the two groups, however, the CI on this estimate is fairly wide (ranging from two months' less progress to one month's additional progress).

As a final sensitivity check we also compute the effect for FSM-eligible pupils from Model 2 (by summing the treatment coefficient and the interaction effect), which is equal to 1.44, which is slightly smaller than the effect estimated in Model 1 but very similar and negligible in size.

In summary, treatment effect analysis from Model 1 suggests that pupils eligible for FSM received no effect from the intervention (made the same amount of progress as their FSM-eligible peers in the control group). This is in contrast to the overall sample that we estimate in the primary outcome analysis to have received a small positive effect. The interaction analysis from Model 2, however, suggests no difference in effect between pupils who were eligible for FSM and those who were not, i.e. that they received a small positive effect as per the primary outcome analysis for the whole sample. There is then a slight discrepancy in the findings (between the two models), but the level of uncertainty around the estimates makes the results inconclusive (and is the likely explanation of this discrepancy). In the efficacy trial, FSM-eligible pupils in the targeted intervention were found to receive a slightly larger effect from the programme as compared to the whole sample.

Analysis in the presence of non-compliance

The purpose of this analysis is to estimate the effect of the programme on pupils who receive a minimum defined amount of the programme ('compliers'). This contrasts with the primary and secondary outcome analysis above, which estimates the average effects for all pupils who were randomly assigned to receive the programme, regardless of how much they

received (an 'ITT' analysis). We define compliance at the pupil level. We consider a pupil a complier in the intervention group if they attended at least 20 Reciprocal Reading sessions. All schools recorded attendance data and sent the registers to FFT at the end of the intervention, who then shared them with BIT. This data suggests that compliance was high with 81% of pupils in the intervention group meeting or exceeding the minimum compliance threshold.

Table 18: Results of the CACE analysis using 2SLS, by compliance definition

Compliance measure	Unadjusted means			Effect size			First stage F-test results	
	Intervention group		Control group					
	n (Compliance = 1)	n (Compliance = 0)	n	Total n (intervention; control)	Hedges' g (95% CI)	P-value	F-statistic F(-10, 3867)	P-value
20 sessions	1,536	371	1,971	3,878 (1,907, 1,971)	0.068 (-0.017, 0.153)	0.115	151.36	< 2.2e-16
18 sessions	1,642	265	1,971	3,878 (1,907, 1,971)	0.064 (-0.016, 0.143)	0.115	256.83	< 2.2e-16
15 sessions	1,778	129	1,971	3,878 (1,907, 1,971)	0.059 (-0.015, 0.132)	0.116	803.48	< 2.2e-16
20 sessions in 12 weeks or less	1,011	896	1,971	3,878 (1,907, 1,971)	0.104 (-0.025, 0.232)	0.115	232.04	< 2.2e-16

To examine the issue of non-compliance, a CACE was estimated. This was done to explore whether low attendance by pupils may be diluting the estimated treatment effect. The results are presented in Table 18 above. Compliance data are available for the full primary analysis sample (n = 3,878). The large F-statistic of the first stage of the instrumental variables model (F(-10, 3867) = 151.36) implies a strong instrument, so the analysis is meaningful.

The estimated effect based on the pre-specified definition of compliance is a Hedges' g of 0.068 (95% CI: -0.017, 0.153) (Table 18), which is qualitatively similar to that of the ITT analysis (Hedges' g = 0.055; 95% CI: -0.014, 0.124) (see Table 15). As with the primary analysis, this is our best estimate of the effect, but there is a reasonable amount of uncertainty in the estimate.

We then ran some additional pre-specified analysis to see how sensitive the CACE estimate is to the definition of compliance. The results where compliance is set at 18 sessions and at 15 sessions are very similar to the result at 20 sessions. This indicates that the model is robust to the threshold.

When the definition of compliance is changed to consider only those schools and pupils that followed the recommended frequency (approximately two sessions per week, for at least 12 weeks), the effect size in Hedges' g increases to 0.104 (95% CI: -0.025, 0.232) (Table 18). This suggests that intensity matters. Pupils who received both the recommended intensity and minimum number of sessions received a larger effect, equivalent to two months progress, as compared to the average pupil in the intervention group who received an effect equivalent to one month of progress (noting that the results from both analyses are consistent with a range from no effect to slightly larger positive effects).

Optimal dosage analysis

In this section, we conduct a descriptive analysis to explore the relationship between the number of sessions received by a pupil and the change in the NGRT overall reading score between baseline and endline.

For a histogram showing the distribution of the number of sessions attended by trial participants, please see the 'Dosage' section of the IPE findings. This shows an approximately normal distribution, with a bit of bunching at zero sessions (which

include those children in the treatment schools that did not do the programme but took the test). The chart suggests that we have enough variability to be able to map the dosage to the change in NGRT scores.

Correlation between the number of sessions taken and individual characteristics

We estimate a linear regression of the number of sessions received on gender, FSM eligibility, EAL status, and baseline NGRT score. We find that FSM eligibility is negatively correlated with the number of sessions received. Being eligible for FSM is associated with doing 1.4 fewer sessions ($p = 0.02$). This could introduce some bias in the optimal dosage analysis, and the direction of the effect of FSM on the change in NGRT score is not clear.

As the outcome is zero inflated, we repeat the analysis using a quasi-Poisson model. FSM remains negatively correlated with the number of sessions. Prior attainment also becomes significant and positive (meaning that higher prior attainment is associated with higher attendance), although the coefficient is negligible in size (equal to 0.0005 and a p-value of 0.01).

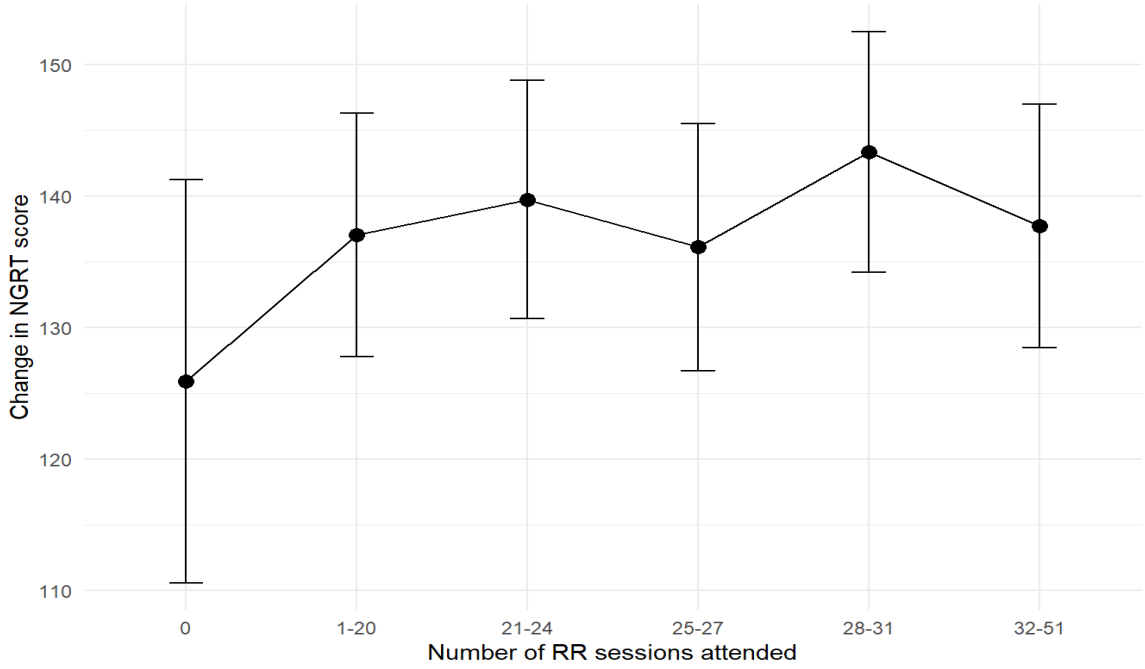
Change in NGRT by number of sessions taken

We divide the distribution of the number of sessions taken in six categories: the reference category is zero sessions, and the rest correspond to the five quintiles:

- Pupils with zero sessions: 214 pupils.
- Quintile 1 (20% of observations) (up to 20 sessions): 412 pupils.
- Quintile 2 (40% of observations) (21–24 sessions): 436 pupils.
- Quintile 3 (60% of observations) (25–27 sessions): 316 pupils.
- Quintile 4 (80% of observations) (28–31 sessions): 367 pupils.
- Quintile 5 (100% of observations) (32 sessions or more): 361 pupils.

We regress the difference between endline and baseline NGRT overall reading scores on the five quintiles and individual pupil characteristics (FSM, gender, EAL status, and baseline NGRT score). Figure 9 presents the estimated coefficients for each quintile, along with 95% CIs. Each coefficient represents the change in NGRT score associated with attending a given number of sessions, relative to attending none, after controlling for pupil demographics and initial reading attainment.

Figure 9: Increase in NGRT overall reading score by dosage



The results suggest that pupils who attended between 28 and 31 Reciprocal Reading sessions saw the largest improvement in NGRT scores. However, the CIs for all quintiles overlap considerably, so we cannot draw firm conclusions about differences in impact based on the number of sessions attended. This lack of precision was anticipated, given the relatively small sample sizes within each quintile and particularly in the reference group (those who attended zero sessions), which limits statistical power.

Nonetheless, the limited evidence available points to 28–31 sessions (approximately 15 weeks of delivery) as a promising target for practitioners aiming to maximise pupil progress—while recognising that this recommendation may be influenced by underlying differences in pupil background (FSM eligibility) and should not be interpreted as causal.

Missing data analysis

Missing covariates

The main 'Outcomes and analysis' section above was conducted using complete cases only. Table 19 below reports the number of observations that had outcome data but were missing at least one covariate, so were excluded from those analyses.

The other covariates in the main analysis (randomisation batch, year group, and month of test) have not been included in the table as they cannot be missing for any pupil by design: year group cannot be missing because all schools were mandated to select two groups of pupils for the programme, one in Year 5 and the other in Year 6, to be randomised; and month of test is always reported in the testing platform together with the NGRT test result.

Table 19 shows that no covariates are missing for more than 5% of the primary outcome sample (or for any of the other outcomes). As a result, we conduct no further missing data analysis on covariates, as the risk for bias was assumed to be very low.

Table 19: Observations with primary and secondary outcome data, by missing covariates

Outcome	Total observations with outcome data	Missing FSM status	Missing gender	Missing EAL status	Missing baseline attainment measure
		n (% of total observations)	n (% of total observations)	n (% of total observations)	n (% of total observations)
NGRT overall reading score / sentence completion score[a]	3,931	0 (0)	0 (0)	0 (0)	53 (1.35%)
NGRT passage comprehension score	3,888	0 (0)	0 (0)	0 (0)	95 (2.44%)
Key Stage 2 SAT reading score	2,091	0 (0)	0 (0)	0 (0)	41 (2%)

[a] *By design, all pupils with an NGRT overall reading score also have a NGRT sentence completion score, so we have reported all missing rates together.*

Missing outcome data

Table 20 shows the missing data rates for the primary and secondary outcomes. Only Year 6 pupils could have Key Stage 2 data at the end of the trial, hence, the smaller sample size.

Table 20: Missing data rates on primary and secondary outcomes

Variable	Group	No. of missing observations	Total observations	Missing rate (%)
NGRT overall reading score / sentence completion score (endline)	Control	144	2,145	6.71
	Treatment	188	2,118	8.88
	Total	332	4,263	7.79
NGRT passage comprehension score (endline)	Control	174	2,145	8.11
	Treatment	201	2,118	9.49
	Total	375	4,263	8.80
Key Stage 2 SAT reading score	Control	26	1,081	2.41
	Treatment	31	1,067	2.91
	Total	57	2,148	2.65

All outcomes from the NGRT test are missing for more than 5% of the sample at randomisation, which could have introduced bias in the treatment effect estimates in the main analysis. The reasons for this missing data have been identified in the 'Attrition' section above of this report.

Next, we try to identify the pattern of missingness for the primary outcome. Given the NGRT secondary outcomes stem from the same test as the primary outcome (as they are subscales), the missing data patterns should be approximately similar, and the outcome data will be missing for the same reasons.

A logistic regression analysis was conducted to determine if individual pupil characteristics (baseline NGRT reading score, gender, FSM eligibility, and EAL status) and treatment group assignment predicted missing primary outcome data. The analysis revealed that being assigned to the treatment group, FSM eligibility, and lower baseline NGRT scores, in descending order of impact, significantly increased the likelihood of missing primary outcome data at the 5% level. This suggests the data is MAR conditional on these variables,[27] though some missingness might still be attributed to unobserved factors. While treatment assignment and baseline attainment are accounted for as covariates in the primary model to control for bias, FSM status, another potential attrition factor, is not. Therefore, a sensitivity analysis was performed by re-estimating the primary outcome model with FSM status as a covariate (Table 21).[28] The results remain largely unchanged, indicating that any bias introduced by factors correlated with the higher absence rate of FSM-eligible pupils in the main model likely has minimal effect.

Table 21: Primary analysis, by sensitivity analysis including FSM status as a covariate

Outcome	Unadjusted means				Effect size			Hedges' g of primary outcome model (95% CI)
	Intervention group		Control group		Total n	Hedges' g (95% CI)	P-value	
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
NGRT overall reading score	1,907 (211)	304.552 (302.605, 306.498)	1,971 (174)	301.687 (299.590, 303.783)	3,878	0.055 (-0.014, 0.124)	0.112	0.055 (-0.013, 0.122)

The logistic regression results indicate differential attrition between treatment and control groups, with the treatment coefficient being statistically significant at the 5% level. This differential attrition occurred because only schools in the treatment group dropped out of the trial. These schools cited resource constraints and staff turnover as their primary reasons for being unable to meet the programme's intensive requirements, which demanded two sessions per week for

[27] Namely, the missingness is related to these variables, not the outcome.
[28] As pre-specified, multiple imputation was not carried out because only the outcome variable is MAR, conditional on covariates. As one of these covariates (FSM status) is not in the main specification, we run this sensitivity analysis that includes it in the model.

each year group. Consequently, two treatment schools completed the endline tests but failed to implement the programme, while three schools withdrew from the evaluation completely.

To address potential bias from this differential attrition, we calculate Lee bounds to estimate an interval for the true treatment effect. The Lee bounds method relies on a monotonicity assumption: differential attrition must flow in only one direction, with no 'defier' units that would have dropped out if assigned to control but remained if assigned to treatment. Since only treatment schools dropped out, we assume that any attrition in the control group resulted from the other individual-level conditional factors previously identified or other unobservable characteristics unrelated to treatment assignment, thus, supporting the monotonicity assumption.

The results of the Lee bounds analysis indicate that, in the absence of bias from differential attrition, the treatment effect could lie between -1.39 and 4.82 NGRT score points. For reference, the treatment effect in terms of the NGRT score of the primary analysis was 2.50.

This wide range reflects uncertainty about what the missing pupils' test scores might have been. To narrow this wide interval, we estimate the Lee bounds conditional on the discrete covariates that determine the attrition (FSM status, NGRT baseline attainment, and treatment status).[29] These trimmed bounds are (-0.88, 4.61).

The Lee bounds provide a range of plausible treatment effects under extreme assumptions about the missing data. Specifically, the lower bound of -0.88 represents the scenario where all pupils who dropped out from treatment schools would have achieved the lowest possible test scores. Conversely, the upper bound of 4.61 represents the scenario where all treatment group dropouts would have achieved the highest possible scores.

Overall, both the basic and trimmed bounds include zero, indicating that we cannot definitively rule out a null treatment effect when accounting for differential attrition. However, the covariate-adjusted bounds narrow from (-1.39, 4.82) to (-0.88, 4.61), with the lower bound moving substantially closer to zero while the upper bound remains relatively stable. This suggests that accounting for pupil characteristics reduces some of the uncertainty introduced by differential attrition. The main study result of 2.50 points falls comfortably within these uncertainty ranges, suggesting that pupil dropouts probably did not dramatically change the conclusions.

The analysis indicates that while losing pupils from the study creates some uncertainty about the true programme effects, the main findings appear reliable.

Additional analyses and robustness checks

Models with no/reduced covariates

As pre-specified, all primary outcome and secondary outcome models were re-estimated without covariates. Additionally, the primary outcome was re-estimated with a model without pupil-level covariates, including only treatment assignment, baseline attainment, and trial design characteristics (the randomisation batch). This model allows for more comparability with other trials by the EEF. All results are presented in Table 22 below.

Table 22: Models with no covariates

Outcome and model	Unadjusted means				Effect size		
	Intervention group		Control group				
	n	Mean	n	Mean	Total n	Hedges' g	P-value

[29] Given that NGRT baseline overall reading score is a continuous variable, we divide the distribution of the baseline attainment in four quartiles and transform it into a categorical variable. Then the sample is divided into cells depending on these four covariates, and the Lee bounds are estimated for each of these cells. Then the group-specific bounds are averaged up, to get the overall effect, weighted by their group's proportion of the overall sample of pupils.

	(missing)	(95% CI)	(missing)	(95% CI)		(95% CI)	
NGRT overall reading score – Only pupil-level covariate is baseline attainment (the EEF standard model, Equation 8 in the Statistical Analysis Plan; Torres Blas and Taylor, 2019)	1,907 (211)	304.552 (302.605, 306.498)	1,971 (174)	301.687 (299.590, 303.783)	3,878	0.060 (-0.006, 0.127)	0.076
NGRT overall reading score – Only treatment covariate	1,930 (188)	304.260 (302.315, 306.203)	2,001 (144)	301.591 (299.490, 303.693)	3,931	0.058 (-0.046, 0.162)	0.272
NGRT passage comprehension score – Only treatment covariate	1,917 (201)	299.447 (297.354, 301.540)	1,971 (174)	297.971 (295.822, 300.119)	3,888	0.031 (-0.070, 0.132)	0.550
NGRT sentence completion score – Only treatment covariate	1,930 (188)	315.650 (313.612, 317.687)	2,001 (144)	313.329 (311.175, 315.484)	3,931	0.049 (-0.043, 0.141)	0.299
Key Stage 2 SAT reading score – Only treatment covariate	1,036 (1,082)	100.904 (100.508, 101.301)	1,055 (1,090)	101.368 (100.995, 101.741)	2,091	-0.073 (-0.213, 0.067)	0.306

For the primary outcome, the point estimates from different model specifications remain almost identical to the main analysis. The model that removes FSM eligibility gives a slightly more precise estimate. The model that includes no covariates gives a less precise estimate, showing that the slight increase in sample size contributes less to the statistical power than the covariates.
For passage comprehension and sentence competition outcomes, removing the covariates leaves the point estimates very similar to the main analysis but reduces the precision.

For the Key Stage 2 SAT reading outcome, removing the covariates leaves the point estimate very similar to the main analysis but slightly increases the precision.

Overall, the results of the main analysis are robust to these changes in model specifications.

Sensitivity checks

Two additional sensitivity checks, that were not pre-specified, were carried out:

1. We estimated the primary outcome model, excluding month of test as a covariate.
2. We estimated the primary outcome model, excluding pupils who experienced IT issues (that may have affected their test performance).

The results are presented in Table 23. Again, there is no substantial difference to the point estimate or CI of the main analysis, so we can be confident that the results are robust to these changes. The variation in time of endline testing does not seem to have introduced any bias. The IT issues experienced by some pupils did not influence the main result.

Table 23: Exploratory sensitivity checks

Outcome and model	Unadjusted means		Effect size
	Intervention group	Control group	

	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Hedges' g (95% CI)	P-value
NGRT overall reading score No month of test	1,907 (211)	304.552 (302.605, 306.498)	1,971 (174)	301.687 (299.590, 303.783)	3,878	0.062 (-0.006 , 0.129)	0.073
NGRT overall reading score No pupils with IT issues	1,900 (218)	304.528 (302.576, 306.481)	1,959 (186)	301.650 (299.542, 303.757)	3,859	0.055 (-0.014, 0.124)	0.117

Estimation of ICC

The ICCs from the study data were as follows.

NGRT overall reading score:

- ICC at the school level, baseline = 0.142.
- ICC at the school level, endline = 0.149.

Key Stage 2 SAT reading scaled score:

- ICC at the school level, endline = 0.273.

Implementation and Process Evaluation results

The IPE research questions we set out to address are listed in the 'Introduction' section of this report. A mixed methods approach was used to address these questions, combining analysis of monitoring and administrative data with observations, surveys, interviews, and in-depth case studies of four schools. This section presents the findings thematically under the following headings: reach; fidelity (dosage, adherence, and quality); programme delivery; pupil engagement and responsiveness; usual practice and programme differentiation; and perceived outcomes, mediators, and moderators.

Reach

Schools

A total of 148 schools delivered Reciprocal Reading. Table 14 in the previous section outlines their characteristics. Among the treatment schools, the average proportion of FSM-eligible pupils was 32.4% (above the national average of 23.8%). The average proportion of Key Stage 2 pupils meeting the expected standard in reading was 72% (almost equivalent to the national average of 73%).

Pupils

Within the recruited schools, the programme aimed to reach Year 5 and Year 6 pupils with good decoding skills but poor comprehension skills, who can read a text accurately but find it difficult to discuss or take meaning from a text and answer questions on it. Schools were responsible for identifying the pupils and were provided with a guide and checklist to help them select pupils who would benefit most from the intervention (see the 'Intervention' subsection in the 'Introduction' section). The guide recommended that schools draw on a range of evidence to identify the target pupils. This was found to be the case for school coordinators and teachers in case study schools who reported selecting pupils based on a combination of recent assessment data and class teacher judgement. In addition, feedback from those who assessed pupils using the checklist provided by FFT suggests that the criteria worked well in supporting judgements—teachers found it straightforward to complete and said it helped them to identify pupils who would benefit.

All schools (intervention and control) were asked to identify potential pupils, and randomisation was carried out after the standardised tests were completed. Baseline assessment data provides some support for the idea that the programme reached the intended pupils. As summarised in the sample characteristics section above, the average overall reading score for our sample is slightly below average for children of the same age (a standardised score of 93.4 vs an average of 100). Another way of expressing this is that slightly more than 60% of the national population will have higher scores than the average pupil in our sample at baseline. When we look at the subsections of the test, we find that the overall reading score in our sample is driven by slightly higher scores in sentence completion and slightly lower scores in passage comprehension, but the sample remains below average for both. This is supported by feedback from the case studies. The teachers we interviewed described the pupils who received the intervention as being secure in their phonics and decoding but weaker in comprehension and sometimes lacking in confidence.

Fidelity (dosage, adherence, and quality)

This section examines the extent to which the training and delivery of Reciprocal Reading sessions were implemented as intended and to the expected quality.

Training

FFT provides two days of training for teachers and teaching assistants who deliver the programme in their school. Day 1 covers the principles behind the intervention, the programme process, resources for delivery, planning for implementation, the opportunity to take part in a Reciprocal Reading lesson at an adult level and observe classroom examples, and guidance and tools for QA. Day 2 focuses on challenging more experienced and improving readers (based on the experience of the pupils on the programme, so far). The training days took place in person, off site during October 2023 and November 2023 (Day 1) and February 2024 (Day 2). As part of the IPE, we conducted semi-structured observations of a purposive sample of six training sessions (four Day 1 sessions and two Day 2 sessions), each delivered by a different trainer.

Fidelity and adaptations

The observations sought to assess how closely the original training plan was adhered to. Overall, the sessions that were observed were largely delivered as intended. Trainers followed the training plan closely. There were only minor differences between sessions, which mainly related to the order of topics or, more exceptionally, the level of detail covered in certain sections.

Teacher engagement

Overall, trainee engagement and participation were high across the Day 1 sessions that were observed. The case study interviews suggest that the trainers' delivery approach and the content of the training contributed significantly to this. Interviewees described the trainers as 'inspiring', 'enthusiastic', and 'knowledgeable' and said that they particularly liked the interactive style of the training. They referenced the fact that there were opportunities to discuss and ask questions as well as experience Reciprocal Reading as a participant. Teachers expressed genuine enthusiasm and described a positive atmosphere in training.

> *Everyone in the room was buzzing, to be honest, like, 'Oh, we're ready to go and teach this now,' whereas some courses, you're like, 'Oh, I've got to go and do this now,' whereas it was exciting.* (Session lead)

However, observers did note some variations in delivery style across the sessions, particularly in the degree of trainer enthusiasm, which may help to explain where engagement and participation from trainees was notably lower (although this was exceptional within the sample).

High levels of engagement were also seen in the two Day 2 sessions that were observed. However, in case study interviews, some participants expressed less positive views about the second training day (discussed further below).

Quality and areas for improvement

Both the training surveys completed by teachers, as well as the observations, suggest a high level of quality across both training sessions.

Teachers' attitudes to both training sessions were positive. This is illustrated below in Table 24, which shows the post-training survey scores across different dimensions of quality. The majority of respondents rated the usefulness of the sessions, the content, and the quality of the trainer highly. The interviews in case study schools highlighted the aspects of the training that school coordinators/teachers found most useful. Modelling was felt to be key.

> *When it was modelled to us, you could kind of imagine how you would do that with your own class.* (Session lead)

Teachers also liked that the training incorporated both theoretical and practical elements, such as covering the rationale for the intervention and providing time to practise and reflect on how it would fit into their school.

Teaching assistants particularly appreciated the clear approach of the trainer who presented the content in a clear and concise way, as well as the mix of roles and experience of trainees, which reduced feelings of nervousness.

> *We both felt out of our depth [beforehand], but actually, it was a mixture of everyone, and the lady [trainer] that delivered it was amazing. She made it quite simple; she made things short; she explained stuff, she was very thorough. (Session lead)*

While the survey ratings for Day 2 remained high overall, the proportion of respondents who rated the second training session as 'very useful' was notably lower compared to the first session (79% and 92%, respectively, Table 24). Ratings were also lower for Day 2 on the structure and quality of the sessions. Qualitative feedback from the case study schools can help to explain this difference. Among school coordinators and session leads interviewed there was the view that the training itself was less valuable second time round. Reasons given included that it repeated things they already knew (some

participants mentioned watching the same videos again), or because it did not meet their expectations for what they would get out of it, such as ideas for how to challenge pupils.

> *It felt to me more of a recap of day one, and I didn't feel like I needed to take time out to go to that.* (School coordinator)

Similarly, respondents rated the quality of the instructor highly across both sessions, however, the proportion who rated the instructor as 'very good' dropped slightly from Day 1 (92%) to Day 2 (84%) (Table 24). The case studies can also provide insight here. Case study participants who were less positive about Day 2 attributed this to the trainer's delivery style, particularly their level of enthusiasm, and ability to explain points in a clear and thorough way.

> *I don't feel like I'm in such safe hands as I did last time with someone who was really clear about what to do, really clear about when to do it, and enthusiastic.* (School coordinator)

This variation in quality did not appear to be linked to the trainer's experience (whether they were already trained in Reciprocal Reading or newly trained up specifically for the trial). These findings suggest some room for improvement in consistency of delivery quality and possibly in the content of Day 2. This room for improvement should however, be considered in the context of overall very high survey approval ratings in terms of quality and usefulness for both training days.

Table 24: Training survey responses (Day 1 and Day 2)

Survey question	Training day	Very useful	Useful	Satisfactory	Not very useful	Not useful at all
How would you rate the usefulness of this training session of Reciprocal Reading?	Day 1[a]	92%	8%	1%	0%	0%
	Day 2[b]	79%	19%	2%	0%	0%
Survey question	**Training day**	**Very good**	**Good**	**Satisfactory**	**Poor**	**Very poor**
How would you rate the structure of the session content?	Day 1	85%	14%	1%	0%	0%
	Day 2	75%	23%	3%	0%	0%
How would you rate the quality of the instructor?	Day 1	92%	8%	1%	0%	0%
	Day 2	84%	15%	1%	0%	0%
Survey question	**Training day**	**About right**	**A bit too fast**	**A bit too slow**	**Much too fast**	**Much too slow**
How would you rate the pace of the session?	Day 1	93%	0%	6%	0%	0%
	Day 2	89%	1%	9%	0%	1%

[a] *N = 357 responses.*
[b] *N = 311 responses.*

The training survey also collected data on respondents self-reported confidence and knowledge as well as actual changes in knowledge through test style questions. For Day 1, there was a large increase in self-reported confidence implementing Reciprocal Reading, from 34% 'very confident'/'moderately confident' at pre-training (128 out of 376 respondents) to 99%

'very confident'/'moderately confident' at post-training (353 out of 357 respondents). Furthermore, the first training session led to increases both in self-reported understanding of Reciprocal Reading from 24% who rated this as 'very good'/'good' at pre-training to 98% at post-training (see Table 26 below) and actual increases in knowledge (see Appendix J Table 1 for more detailed results of the pre- and post-regression analysis conducted on these items).

The increase in confidence and knowledge shown in the survey results was also reflected in the qualitative feedback. Participants said they came away from the initial training knowing what to do and with the confidence to deliver, which was attributed to the thoroughness of the training and the knowledge of the trainer. This was also evident in the observations. In the most high-quality sessions, observers saw trainers making links between different topics and bringing in their own personal experience, which is likely to have contributed to this growth in confidence and knowledge.

Table 25: Changes in confidence implementing Reciprocal Reading before and after Day 1 training

	Very confident	Moderately confident	Neutral	Moderately unconfident	Very unconfident
Please rate how confident you would feel implementing the Reciprocal Reading programme with pupils at your school	Pre-training				
	7%	27%	26%	20%	20%
	Post-training				
	59%	40%	0%	0%	1%

Table 26: Changes in self-reported knowledge of Reciprocal Reading before and after Day 1 training

	Very good	Good	Satisfactory	Poor	Very poor
Please rate your understanding of the skills, knowledge, and attributes needed to effectively deliver the Reciprocal Reading programme	Pre-training				
	5%	19%	34%	27%	14%
	Post-training				
	64%	34%	2%	0%	0%

The main aim of the second training day was to upskill teachers to ensure they can effectively engage, and challenge experienced and improving readers. The survey results indicate that the training helped teachers to develop their understanding of how to do this. The increase was less stark than that seen in the pre- and post-training scores for Day 1, as most teachers (83%) already rated their understanding as 'very good'/'good' (Table 27). Nonetheless, this increased to 99% post-training, with the biggest increase being the proportion of teachers who felt their understanding was 'very good' (64%) (Table 27).

Table 27: Changes in understanding of how to engage and challenge experienced pupils before and after Day 2 training

	Very good	Good	Satisfactory	Poor	Very poor
Please rate your understanding of the skills, knowledge, and attributes needed to effectively engage and challenge experienced pupils in the Reciprocal Reading programme	Pre-training				
	20%	63%	15%	2%	0%
	Post-training				
	64%	35%	2%	0%	0%

Overall, the observations, training feedback survey, and case studies suggest that the training was well delivered and received. The growth in the pool of trainers from the efficacy to effectiveness trial (five new trainers) did not appear to have any substantial implications for the quality of the training.

Programme delivery

Overall, the intervention was delivered with a high degree of fidelity. As the subsections below explain, there was high compliance with the minimum dosage requirements, schools adhered to the session structure as far as possible and delivery was judged to be of good quality.

Dosage

Treatment schools (n = 148) were asked to deliver two weekly 20–30-minute sessions over 12 weeks to a small group of approximately six pupils (a minimum of four pupils and a maximum of eight pupils). Monitoring data was collected to report on the extent to which schools adhered to these requirements.

Table 28 shows the average group size and number of weeks and sessions delivered by treatment schools. The average number of pupils per group was seven, with the minimum reported as five and the maximum as eight. Out of 148 schools, 125 schools reported delivering at least 12 weeks to Year 5, and 131 schools said they had delivered at least 12 weeks to Year 6. In total, schools delivered an average of 28 sessions across 14–15 weeks (an average of 1.8 sessions per week). This was substantially lower than the average dosage reported in the efficacy trial for the targeted intervention (approximately two sessions per week for 26 weeks).[30] This partly reflects the fact that a larger number of schools took part in the effectiveness trial, which resulted in extended timescales for training and testing and a shorter implementation period. Furthermore, the CACE analysis found that doing 20 sessions in 12 weeks or less resulted in a bigger treatment effect. The lower overall dosage and fact that sessions were delivered over a longer time span may help to explain why the impact evaluation found only a small positive effect on pupils' overall reading score and a negligible impact on the secondary outcomes (reading comprehension, sentence completion, and Key Stage 2 SAT reading score).

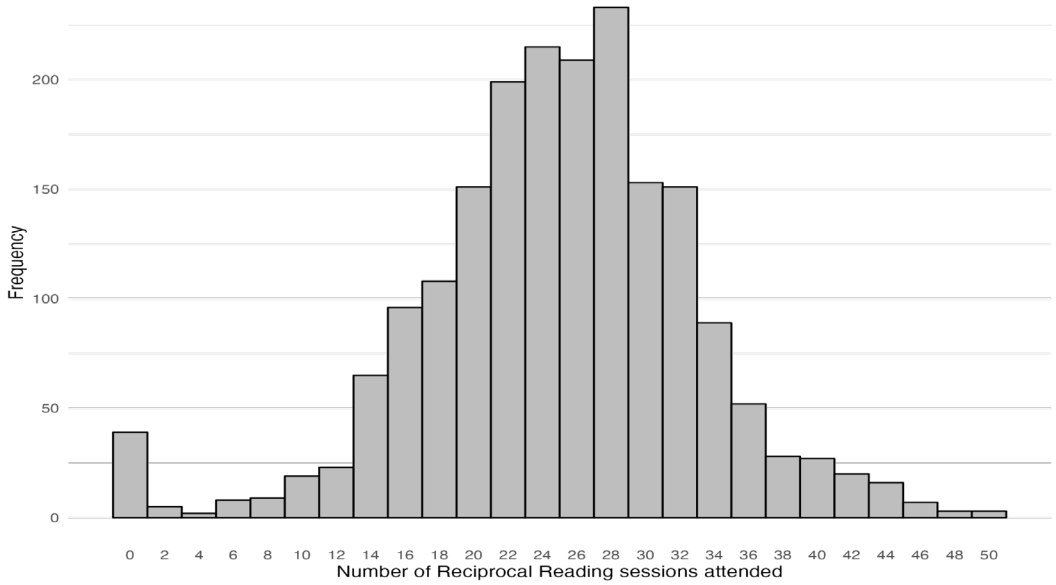
There was some variation in delivery, shown by the minimum and maximum number of sessions. The higher-than-average maximums (62 and 51 for Year 5 and Year 6, respectively) can be explained by the fact that some schools delivered sessions more frequently, in some cases on a daily basis (Table 28). There were also two treatment schools who did not deliver the programme but who were included in the impact evaluation, which can explain the lower minimum range (this is also reflected in the pupil attendance statistics shown in Table 29 and Figure 10 below).

Table 28: Dosage (how much of the intervention was delivered to pupils in the treatment group)

	Sessions delivered		Weeks of delivery		No. of pupils per group	
	Year 5	Year 6	Year 5	Year 6	Year 5	Year 6
Mean	28 (rounded)	28 (rounded)	14.8	15.3	7	7
Minimum	1	1	1	1	5	5
Maximum	62	51	24	24	8	8
N	148	148	148	148	148	148

[30] In the efficacy trial, dosage was defined as the time spent delivering the programme to pupils. The mean total number of minutes of delivery of the targeted intervention was 1,707 minutes (65.77 minutes per week). This equates to approximately two sessions per week, assuming a 30-minute session length (the recommended duration and the average observed in this present effectiveness trial). (O'Hare *et al.*, 2019).

Figure 10: Distribution of sessions



Data on the length of sessions was provided by schools who responded to the retrospective survey. Sessions lasted between 15 minutes and 45 minutes (only three respondents said that they lasted 15 minutes). The average session length was 30 minutes (this is the recommended duration, although schools are permitted to deliver more than this). The sessions we observed in case study schools ran for approximately 30–45 minutes. In the interviews, session leads noted that sessions lasted longer where additional support and explanation was needed, or where pupils were highly engaged with the story. Teachers also noted that the length of sessions had changed over time, with sessions becoming shorter as pupils became more familiar with the structure and the strategies (which is expected by the developers).

Table 29 shows the number of sessions and weeks attended by all pupils who received the programme and sat the final NGRT exam (and so are part of the impact evaluation). The monitoring data show that out of those 1,930 treatment school pupils, 1,552 (80%) completed at least 20 sessions, and 1,574 (82%) completed at least 12 weeks.[31] Figure 10 shows the distribution of sessions attended. As mentioned above, this data includes pupils from two schools who took the test but did not attend any sessions as their school did not deliver the intervention. In schools that did deliver the intervention and did not withdraw from the trial, pupil attendance may explain why not all pupils attended a minimum of 20 sessions/12 weeks. In the case studies, teachers said that this was a wider problem, especially for pupils from less stable backgrounds, such as children from the travelling community. This is reflected in the quantitative analysis, which found that FSM eligibility was correlated with doing fewer sessions.

Table 29: Pupil attendance

	Sessions attended		Weeks attended	
	In total	Before the endline test	In total	Before the endline test
Mean	25.4	25[a]	14.3	14.1
Minimum	0	0	0	0
Maximum	51	51	24	23
N	1,930[b]	1,930	1,930	1,930

[a] *This average includes pupils that did zero sessions but only includes those that did the endline test.*

[31] Pupils were considered compliant if they attended a minimum of 20 sessions.

[b] *N = 2,106 includes all pupils in the treatment group whose parents did not opt them out of the evaluation. Out of those, N = 1,930 represent the pupils who also did the endline test.*

Nevertheless, treatment school pupils who completed the endline test (and were included in the impact evaluation) attended an average of 25 sessions over 14 weeks, exceeding the minimum dosage specified by the developers. This is slightly less than the optimal dosage estimated in the exploratory analysis above, which suggests that pupils who attended between 28 and 31 sessions (approximately 15 weeks of delivery) made the largest improvement. So, a slightly low dosage could be part of the explanation for the small treatment effect.

The IPE identified three main school-level factors that acted as barriers or enablers to delivering the required dosage.

First, schools had to fit the sessions into the existing school day. This was challenging for some—65% of respondents to the retrospective survey selected 'timetabling issues' as a barrier to implementation. In case study schools, teachers said that SATs made it particularly difficult to find space in the timetable for Reciprocal Reading for Year 6. Nevertheless, the statistics on dosage suggest that most schools found a way round this. Adopting a flexible approach to delivery (e.g. rearranging sessions to take place on another day during the week if needed) and delivering the intervention during time protected for independent/guided reading both helped. The latter removed the need for staff to find an additional time slot for Reciprocal Reading during the school day.

Second, the intervention required staff to be released from normal duties to deliver the intervention on a weekly basis. Insufficient staff resources was a constraint for some participating schools. This was cited as a challenge to implementation by 32% of respondents to the retrospective survey. This was also linked to timetabling as teachers said that it could be difficult to do two sessions a week when cover was needed elsewhere due to being short staffed.

> *It's mostly been okay, but there have been weeks where it hasn't been able to happen. Like me, for example, today we're four staff down already because of swimming and someone's just gone home sick, so there are times where I'm pulled here, there and everywhere. So, it has been hard, I have to say, to do sometimes, to make sure it's done twice a week.* (School coordinator)

The high level of fidelity observed in the face of such barriers is possibly a further indicator of the value that teachers saw in the programme. Lack of staff resources, however, was the reason why one treatment school withdrew from the trial and did not deliver Reciprocal Reading. Among schools who were able to deliver, strong senior buy-in was important as it helped to ensure staff were available for the intervention.

> *[The principal] was really involved in it, and really on board with it. Again, was like, 'We need to make cover happen to release these staff, to get them on Reciprocal Reading.' So, that was really, really beneficial and supportive.* (School coordinator)

The fact that the intervention can be delivered by teaching assistants was also important—out of the 179 session leads who completed the retrospective survey, 129 were teaching assistants. This suggests that many treatment schools were able to use teaching assistants to deliver the sessions.

Third, school infrastructure/space was of lesser importance to survey respondents, but nevertheless, 29% of respondents indicated that lack of physical space to carry out the sessions was a challenge for implementation.

The IPE also examined variations in dosage by region (using regional clusters specified by FFT, based on the location where the school staff attended the first training day).[32] The areas with the most variability in the number of sessions delivered were Birmingham, closely followed by the online training cluster. Bristol, Birmingham, and online schools did the highest

[32] Note: the statistics were obtained from the number of sessions delivered by each school in total. This adds all sessions for the two groups (Year 5 and Year 6). The trainer was the same for all sessions delivered in a region, with the exception of London, which had three different trainers in Training Day 1.

average number of sessions, while Bedford, Sheffield, and the Black Country had the lowest averages, and did 15 to 17 sessions with each group on average. The only relevant covariate we have in the data to try to explain these regional differences is an indicator of trainer experience. We see no relationship between trainer experience and dosage. The developers also saw no explanation for these differences.

Adherence

As well as dosage, the IPE explored the extent to which delivery of Reciprocal Reading adhered to the intended model. Overall, there were high levels of adherence to the model (this is explored further below in relation to the use of appropriate materials and session structure). This was underpinned by teachers' attitudes towards the intervention, which were overwhelmingly positive—the programme was described by over three-quarters (75%) of treatment school respondents as 'a lot'/'somewhat easier' to implement compared to other reading interventions. Furthermore, a prominent theme that emerged from the case studies was teacher buy-in and confidence in the effectiveness of the intervention, which had been instilled through the training.

> *It does feel like something that is going to work, which is why it's been easy to buy into. It's been a very straightforward process.* (School coordinator)

Use of programme materials

Schools received resources from FFT to help them deliver the intervention. Evidence from the IPE suggests that schools used the materials as intended. Almost all respondents in the retrospective survey (99%) said that they had followed the materials provided by FFT when planning the sessions, the majority of whom (79%) said that they closely followed the materials. This is supported by the qualitative feedback gathered from the case studies. Teachers reported delivering sessions according to the example lesson plans, particularly at the outset, and using the templates/crib sheets for their own planning, which in turn reduced their workload. The materials were perceived to be of high quality and easily accessible.

Use of appropriate text and dictionaries

The sessions we observed in case study schools used texts from the anthology of short stories provided as part of the programme materials. Teachers felt comfortable using the anthology knowing that these were specifically selected for the intervention. There were, however, varying levels of confidence among teachers about selecting texts independently. Some expressed concerns about choosing a text with the right level of challenge or selecting a different genre (e.g. non-fiction). This suggests that teachers would benefit from more support/guidance on choosing and applying Reciprocal Reading to different texts, which could otherwise be a barrier to intervention sustainability.

The programme handbook specifies that all pupils should have access to a dictionary. Schools were provided with dictionaries in the resource pack. Among the case study schools, dictionary use varied. Where dictionaries were not used, this was a deliberate decision taken by the session lead who felt that pupils did not have the skills to use it effectively, and felt other ways of clarifying vocabulary (e.g. working out the meaning through discussion) were more useful.

> *So we tried with the dictionaries to start off with, but I felt like it was really, really slowing down and taking away from the enjoyment of it. Just for the fact that I think using a dictionary is a skill in itself. Unless they're really, really secure in using the dictionary, then it becomes a bit of a, 'Oh, I've got to go and find this word in the dictionary.' Then they'd find the word and not necessarily know how to put it into the context of it, anyway. (Session lead)*

Four-strategies structure over three cycles

The intervention consists of pupils reading a text using four reading strategies (predict, clarify, question, and summarise), which are repeated three times per session (referred to as rounds or cycles). On average, respondents to the retrospective survey reported completing two to three rounds per session (slightly lower than the recommendation of three to four rounds). The length of sessions reported in the retrospective survey ranged from 15 minutes to 45 minutes, with the average session length reported as 30 minutes (the recommended duration.) Fitting the three rounds into this time slot was raised

as a challenge. Teachers said that, on the occasions where they only had time for two rounds, this was due to a rich discussion within the group, which they were reluctant to disrupt.

> *If you say no, there must be three rounds in every session, you lose some of the magic of it. Actually, they were saying the reason they're only doing two rounds is because there is so much conversation, and that's not something I want to stunt.* (School coordinator)

This points to high levels of pupil engagement (discussed further below) but suggests that time constraints and issues of time management were barriers to adhering to the three cycle, four-strategies structure.

The groups we observed followed the four-strategies structure over three cycles, however, the following observations were noted:

- **The third cycle tended to be shorter and less discussion-based.** Time constraints (the session coming to an end) appeared to play a part in this. However, in the interviews, session leads also said that they deliberately aimed for the second and third rounds to be quicker and more pupil-led, after modelling the first round quite heavily.

- **Pupils' responsiveness to the individual strategies varied.** Most notably, the pupils in the sessions that were observed struggled to construct summaries, sometimes giving a prediction instead, or leaving out key details. Across the case studies, teachers confirmed that pupils found summarising the most challenging of the strategies and said it was generally a hard concept to grasp. In response, some session leads had introduced paired work (asking pupils to come up with three summary points with their partner) and sentence starters.

Adaptations

Schools were free to make certain delivery adaptations. The retrospective survey asked respondents what adaptations or changes they made. Adaptations reported in the survey included running sessions more frequently (17%) and using electronic devices (16%). In the observed sessions, iPads™ were used for clarifying vocabulary and served the purpose of a dictionary.

The observations and interviews also identified a range of adaptations related to pupil needs. Across the sessions that were observed, Session leads were attentive to individual pupils' needs. For example, teachers noticed where individual pupils were struggling, and offered support either by modelling themselves or asking others to model. During the interviews, session leads also described adaptations such as providing dyslexic pupils with coloured overlays and considering individual needs of pupils when choosing words to clarify.

As set out in the intervention description, teachers and teaching assistants using the reading strategies in other subjects (embeddedness) was also considered as an acceptable adaptation. While this was not a direct modification to delivery, we discuss this point here. The level of embeddedness varied across the case study schools. There was evidence that some schools had taken additional steps to embed the intervention by providing internal training to other members of staff and trialling the strategies with other year groups. There were several reasons for this. First, participants said that they believed the programme was a 'whole-school responsibility', and that whenever they introduced something new, they would try to make all staff aware. Second, they were motivated to introduce the strategies in other contexts due to interest in using the approach beyond the trial. Pupils in these schools are likely to have been exposed more to the strategies outside of the two sessions a week. In other schools, there had not been a deliberate decision to introduce the strategies outside of the treatment group but nevertheless teachers had observed pupils referring to them.

Quality: Facilitation style and skills

The observations also sought to assess how well teachers/teaching assistants delivered the programme. A set of quality indicators were developed, with input from FFT, and used in the observations of Reciprocal Reading sessions. These are discussed in turn below.

1. **Ability to engage the group in discussion and nurture pupils' self-concept as a reader.** Session leads were able to engage pupils well, for example, by relating the story to their own lives and asking them how they would feel and showing interest in pupils' predictions about the story. Across the observed sessions, teachers and teaching assistants were supportive and encouraging. For example, session leads thanked pupils for their contribution, praised them for working out meanings, let them know if they were on the right lines, and provided reassurance when they missed something in the text.
2. **Responsiveness to pupil interests and needs.** The training provided by FFT emphasises the importance of facilitators responding to lines of enquiry raised by pupils during the session, and if appropriate, deviating from the plan to allow for this. This was evident where session leads allowed pupils to discuss points of interest. For example, in one of the sessions the facilitator allowed pupils to have a discussion about the main object in the story—a balaclava—including what it was and when it might be worn. However, this was sometimes more limited in order to keep sessions on track.
3. **Pupils rely less on the teacher for direction as cycles progress.** This indicator was less visible in the sessions we observed. Instead, we saw that pupil reliance on the teacher fluctuated according to the strategy. As discussed in the previous section, pupils were generally able to take more ownership over their predictions, whereas session leads had a more active role in prompting pupils to identify vocabulary to clarify and questions to ask. Pupils also needed more support constructing summaries.

Several factors influenced the quality of the sessions:

- **Staff experience and skills.** School coordinators who were interviewed were confident that the staff who were delivering had the skills and experience to deliver to a high standard. The fact that school coordinators trusted their staff is also likely to have had a positive impact on their performance and the quality of delivery.

- **Training and resources.** In the interviews, teachers reported that the resources provided by FFT were easily accessible and facilitated planning.

  It was so straightforward and the fact that there were plans in place, so you didn't have to go in cold trying to find your own words, trying to think of your own questions. (Session lead)

- **Planning.** It was evident that session leads had planned out the sessions that were observed. The most prominent example of this was where teachers had identified words that may need clarifying and offered their own predictions/questions, which they had prepared in advance.

- **Time.** Teachers commented that they sometimes felt restricted by time. They found it difficult to get through everything in the recommended time but were not always able to extend the sessions to allow for this. As stated above, this is likely to have influenced how they ran the sessions to ensure they got through the cycles.

  At times, it was hard to fit the sessions in as the group was very engaged and sessions would last up to 45/50 minutes. (Teacher, Retrospective survey)

Pupil engagement and responsiveness

Pupil engagement with the intervention was high. Around 98% of retrospective survey respondents said that pupils had been 'very engaged'/'somewhat engaged'. This was also reflected in the case studies. Overall, pupils were highly engaged in the sessions that were observed, albeit with some differences between pupils. Four factors were identified as influencing pupil engagement: i) the nature of the programme; ii) the stories; iii) pupil attitudes towards school; and iv) group dynamics.

First, teachers attributed high levels of engagement to the nature of the intervention—the interactive style and use of discussion meant that pupils did not see it as 'reading', which they typically associated with reading independently and written questions.

> *If you'd have come in and compared an old, guided reading style lesson to that, then it would have been totally different. Like I said before, a lot of them were like, 'Oh, it gets me out of reading,' but they read more in them sessions than they do in any other point of the school day. That's the enjoyment of it, that they don't actually see it as reading, which, for a lot of kids aged ten and eleven, has a lot of negative connotations to it.* (School coordinator)

This was also reflected in the pupil focus groups.

> *Where everybody else is reading, we're having these fun sessions.* (Pupil)

Overall, the anthology stories were received positively by pupils. This was reflected in the observations, teacher interviews, and pupil focus groups. Teachers said that pupils were generally interested in the stories and enjoyed reading them, although they noticed some variation with some texts less engaging than others. Similarly, across the case studies, pupils were generally positive and enthusiastic about the stories, were able to recall specific storylines, and were clearly looking forward to finding out what was going to happen next. The length of the stories played a part in this—pupils liked being able to finish the story, which some teachers said was often not possible in usual guided reading (often focused on a chapter or section of a text). According to session leads, where the stories lacked suspense, this could affect pupils' engagement and enjoyment.

There were some differences in engagement between pupils, which was evident in body language (e.g. not following along with the text, or closing the book). Session leads said that lower engagement levels were generally linked to general attitudes towards school/learning rather than the intervention. There were some instances of disruptive behaviour, but these were managed well (see below) and did not affect the quality of the sessions. Where teachers were having to manage disruptive behaviour, they thought it might help to have a smaller group of pupils so they could pay more attention to these pupils (though this would obviously increase the burden on staffing).

Group dynamics were also seen to influence engagement, particularly participation in the discussion. Across the observed sessions some pupils were less vocal and others more dominant. Teachers managed different levels of engagement by prompting pupils to contribute where they noticed issues with concentration and reminding dominant/vocal pupils that they should allow others to speak. In one of the groups where boys were dominating the discussion, the teacher went around the group and asked each pupil to share a prediction and summary.

Pupil responsiveness to the intervention, and specifically the four strategies, varied. This is discussed further in the subsection 'Moderators' below.

Usual practice and programme differentiation

The IPE captured usual practice in treatment schools and the extent to which Reciprocal Reading was perceived to differ (or not) from business as usual.

In the case studies, teachers said that their business as usual typically consisted of 'guided' or independent reading. The format of these sessions varied but ranged from pupils reading independently and answering questions to small group work. Teachers mentioned running booster/support groups for Year 6 pupils geared towards preparing pupils for SATs.

When asked what they thought the distinguishing features of the intervention were, compared to other types of reading support, teachers highlighted the following:

- **Flexibility.** Teachers felt it was less prescriptive and rigid than they expected:

  > *It was very clear that there wasn't one way to do it because the person we had leading our training said, 'You can do some of the reading if you want to. It doesn't have to be completely independent if they're not ready for that.' I felt that we can adapt it.* (Session lead)

- **Strategies.** These provided a scaffolded approach to reading.

- **The discussion element.** Teachers said this made it 'low stakes' and different from every other class/activity children are doing:

  The fact that they don't have to write, I think takes a lot of pressure off, and so they can just concentrate on the discussion and concentrate on their thinking and their reflecting. (School coordinator)

The latter features were felt to be particularly unique to the intervention. Teachers said that they would usually teach these strategies separately, and guided reading would typically involve written work. Difficulties teaching comprehension effectively was one of the reasons why schools signed up to the trial, further highlighting the role of Reciprocal Reading in filling a gap.

The survey data on usual practice shows that pupils in the treatment group had access to differing levels of support outside of the intervention. In terms of other reading interventions offered by treatment schools, 37% of respondents indicated that Reciprocal Reading substituted an alternative intervention, 14% of whom said it completely substituted an alternative. Approximately half (49%) of respondents said it did not replace any interventions. The retrospective survey also asked treatment schools whether they were using concurrent interventions for Key Stage 2 pupils during the trial. The results show a mixed picture—43% of respondents said that their school offered other interventions focused on reading comprehension at the same time as Reciprocal Reading, while 56% did not.

Control schools who completed the retrospective survey reported which activities their school used to improve Key Stage 2 pupils' reading skills. The most common activities were teachers reading aloud to pupils (98%), independent reading time (94%), and pupils completing reading assignments during lessons (92%). Control schools were also asked about any additional support that was given to the pupils selected for the intervention, to explore risk of contamination. Around 70% of respondents said that the pupils that were selected received additional, targeted support to help develop their reading skills. The types of additional targeted support they received is reported in Table 30. Group sessions were the second most popular option. The results suggest this support was delivered regularly—67% of respondents said that selected pupils received this support at least once a week.

Table 30: Additional support that was given to the selected pupils in control schools

Additional support	%	n
Catch-up group sessions	67%	64
Catch-up individual sessions	21%	20
Targeted in-class support	78%	74
Pre-teaching	36%	34
Pupils completing reading assignments during lesson time	36%	34
Pupils completing reading assignments at home	6%	6

The evidence from the control school survey suggests that, on the whole, most control schools did not change their practice during the trial. When asked whether the pupils in the control group would have received the additional, targeted support regardless of their selection for Reciprocal Reading 97% answered 'yes'. However, a lot of targeted reading support of varying kinds was clearly delivered to pupils in the control group. A total of 38 schools in the control group used reading interventions from other organisations including PiXL (Partners in Excellence), FFT (Tutoring with the Lightning Squad), and

Reading Plus.[33] This may help to explain why a larger difference was not seen between the treatment and control groups. The efficacy trial captured information on the school comprehension ethos and reading strategy behaviours in control schools but not details on the types of targeted support/interventions that were delivered, which limits our ability to make comparisons around the level of targeted support control pupils in the two trials received.[34]

Furthermore, from the retrospective survey for control schools there was some evidence of non-compliance, which may have biased the treatment effect. A total of 17 schools in the control group reported occasionally using elements of Reciprocal Reading in their usual teaching practice or their targeted support for the 12–16 selected pupils. Three schools said they fully integrated the programme in their teaching practice or support.[35] It is unclear from the survey what led to this contamination in the control group. The survey did not include questions about teachers' previous experience of being trained in or using the Reciprocal Reading approach. Furthermore, we did not carry out any qualitative research in control schools and the two schools who withdrew from the treatment group after receiving the training said that they did not deliver Reciprocal Reading in their responses to the control survey. However, we know that FFT had trained more than 5,000 teachers in the approach at the point of the trial that, reciprocal teaching as an idea has been around for a long time, and that some resources on approaches that are similar to the Reciprocal Reading programme are freely available online. So, it is clearly possible that some teachers in control schools had been previously exposed to Reciprocal Reading and related approaches and may have been formally trained in it.

Perceived outcomes, mediators, and moderators

This section presents evidence from the IPE on perceived outcomes of the intervention and views on how and why these outcomes arose.

Reading ability and comprehension

Findings from the retrospective survey suggest that Reciprocal Reading is perceived to improve pupils' reading proficiency. Most respondents (82%) believed that, compared to their school's usual teaching, Reciprocal Reading was more effective at boosting reading skills, with just under a quarter (24%) saying it is much more effective. This is supported by findings from the case studies—there were teachers who had recent assessment data available to them and said that this was showing positive results, including for pupils who were not receiving any additional reading support outside of classroom teaching and Reciprocal Reading.

Survey respondents said the most important factor driving this improvement was the fact that the intervention gave pupils a more structured approach to reading through the strategies. This supports the theory that teaching pupils the four strategies helps them to better comprehend what they are reading. This was also highlighted in the case studies—teachers in particular, attributed progress to the fact that the intervention taught pupils the core skills needed for reading in an integrated way (through the cyclical aspect).

> *I'd like to continue [Reciprocal Reading] because when you break it up it's easier to remember what's happening in the story. Because when I read a book in the classroom, I just forget about it the second I've finished with the book.* (Pupil)

[33] PiXL offers schools a range of resources, assessment tools, and targeted strategies. Their approach is underpinned by a model of 'Diagnosis, Therapy, Testing, and Re-visit (DTTR)', which encourages schools to identify specific areas of weakness, implement targeted support, assess progress, and then continually review. Tutoring with the Lightning Squad is a structured intervention aimed at pupils in Years 1 to 6 who are reading below their expected level. The programme uses a blended learning model that combines face-to-face tutoring with an online platform of activities. Reading Plus is an adaptive online literacy tool, designed to develop reading fluency, comprehension, and vocabulary.

[34] The results showed that the use of small group activities to teach reading comprehension was similar in both groups at pre-testing. Around 76%/78% of headteachers in control/intervention schools, respectively agreed that current teaching in Years 4, 5, and 6 makes use of small group activities for the teaching of reading comprehension.

[35] In the efficacy trial, control schools were not asked directly whether they had used any elements of Reciprocal Reading, so we are unable to compare control group activity between the two trials.

I think it's slowly drip-feeding that kind of understanding of what you should be doing yourself as a reader, rather than just answering loads of questions. (Session lead)

There was also evidence that the strategies had given pupils a framework to apply to texts more widely (outside of the sessions). In some cases, teachers said their pupils had used the strategies and approach (breaking the text up into manageable sections in order to discuss and monitor understanding) during practice SAT papers (though the impact analysis does not find a positive effect for Year 6 pupils on Key Stage 2 SAT reading scores). This is significant as use of the strategies is considered a mediator that links the activities and outcomes of the intervention. Together these findings can be seen as evidence of some of the core programme mechanisms and can help to explain the positive effect on overall reading score. However, while staff in case study schools were largely optimistic about the difference Reciprocal Reading was making to pupils reading comprehension, some teachers felt that this was difficult to assess without data. Others had seen progress but were uncertain about how much they could attribute this to Reciprocal Reading.

Similarly, among the parents who were interviewed, some said they had noticed improvements in their children's reading abilities over the course of the academic year, though they said they did not know if this was attributable to the programme as their awareness of it was very limited. Echoing teachers, they also pointed out reading comprehension as an aspect that had significantly improved.

Comprehension was an issue. He picked up the words but not the story. That has changed. With the comprehension and confidence, I feel like the progress has definitely been in parallel with the programme. He doesn't just read the words anymore. (Parent)

Confidence/attitudes towards reading

According to the programme logic model, changes to pupils' reading behaviours is one of the ways in which the intervention is expected to improve reading attainment. While it is not possible to comment on frequency of reading (e.g. whether pupils were reading more for pleasure in their own time), the IPE suggests that there were changes in how pupils felt towards reading, particularly their confidence. Around 98% of retrospective survey respondents said that Reciprocal Reading had improved their pupils' confidence. Teachers said that this was illustrated by their willingness to actively participate during the sessions.

I think what you've seen over time is, actually, all of the children are now happy to participate. Happy to predict. Putting their hands up, giving answers. So, I think, in terms of the confidence, for the vast majority, it has grown. (Session lead)

This was also reflected in the pupil focus groups. Pupils said that they felt more confident and less embarrassed speaking in the sessions than they did at the start of the programme, especially about vocabulary they did not understand.

During the interviews with parents, participants gave examples of where they had seen changes in their child's reading habits, which reflected an increase in their confidence and enjoyment of reading. In particular, children being more proactive about reading and choosing to do so independently was a recurring theme. Some parents also cited changes in the way their children involve them in their reading. Reading aloud for them, narrating the stories to them of their own accord after having read, or asking them for help with the meaning of a particular word, stood out as changes frequently mentioned by parents that are also reminiscent of the Reciprocal Reading strategies, suggesting a possible connection between the programme and these improvements.

Perceived outcomes for specific groups

Disadvantaged pupils (FSM-eligible)

Most respondents to the retrospective survey felt that Reciprocal Reading had a positive impact on their FSM-eligible pupils. When asked how big a role they thought Reciprocal Reading played in their FSM-eligible pupils' reading progress this school year, more than half (60%) said it somewhat boosted their progress, with an additional 17% saying that it significantly

boosted their progress. The top-rated reasons for this were: that it is more interactive (78%); that it gives them a more structured approach to reading (73%); and that it is carried out in small groups (67%).

As part of the IPE, we sought to investigate the hypothesis that Reciprocal Reading can have a levelling effect on inequalities in the home reading environment by providing disadvantaged pupils with a stimulating group reading environment, and techniques to comprehend texts more effectively. There is evidence from the case studies that can support this hypothesis. Teachers highlighted specific aspects of the intervention, which they thought may be more impactful for disadvantaged pupils who may have a limited home reading environment and lower levels of literacy.

For example, where teachers felt they had seen the most improvements in disadvantaged pupils, they linked this to the fact that Reciprocal Reading ensured they were having an opportunity to read with a teacher, be listened to, and discuss a text, which they may not otherwise have at home. Others highlighted that Reciprocal Reading was unique in its ability to level the playing field for pupils with lower levels of literacy. They said that having multiple cycles in a session allowed teachers to pick up on skills that pupils might be struggling with and tackle this immediately.

The subgroup analysis reported in the impact findings above does not really illuminate this discussion. It suggests that pupils eligible for FSM either received no effect or a similar effect to all pupils (depending on the type of analysis), but there is a fairly high level of uncertainty around all estimates that probably explains the discrepancy in results.

EAL pupils

A large proportion of respondents to the retrospective survey agreed that Reciprocal Reading had a positive impact on their EAL pupils. Around 85% of respondents said that Reciprocal Reading had a large or moderate increase on their EAL pupils' confidence. Around 83% believed that Reciprocal Reading had boosted their FSM EAL pupils' reading progress. They said this was because the programme helped them to learn new words (83%), was interactive (78%), and provided a structured approach to reading (73%). This was reflected in the interviews with teachers who said that EAL pupils in particular were more confident and engaged and had responded well to the structure of the sessions.

> *Knowing, first we do this, and then we do this, and then we do this. So, that's really supported them. Whereas before, because it was so heavily based on comprehension style questions, if they didn't understand the text, they just got nothing on the questions.* (Session lead)

Moderators

We identified four main factors that moderated the effects of the intervention. Some of these factors are discussed in detail elsewhere in the report, in these cases we signpost to these sections.

Pupil and school characteristics

Certain characteristics of the schools and pupils who took part in the intervention emerged as important moderating factors through the IPE. Around 27.1% of the treatment school sample were EAL pupils (see Table 13). The evidence from the teacher interviews indicates that having EAL may influence outcomes for pupils. On the one hand, teachers observed that EAL pupils could face additional barriers accessing the intervention, due to difficulties understanding specific idioms and phrases, and said that this was a wider challenge in their school for pupils who do not speak/read in English at home.

> *The biggest area of need in this school is communication and language, which is why we have got so many reading-based or phonic-based interventions. Our children all come in very much with English as a second language. Eighty-two per cent of our children have English as a second language.* (School coordinator)

Nevertheless, there was also the view that EAL pupils had especially benefited from the intervention because of the focus on clarifying vocabulary. However, as this only relates to one of the four strategies it may not be sufficient to have an impact on overall reading proficiency. It is therefore, unclear as to how beneficial the programme is for EAL pupils, and they represented too small a group in the sample for a subgroup analysis.

FSM status and prior reading skills are also likely to have moderated the effects of the intervention and may be interlinked. The case studies were sampled purposively to include three schools with a higher-than-average proportion of FSM-eligible pupils. Among these schools, teachers told us that reading was a priority due to it being a struggle for many of their children, which was often linked to children entering the school with low literacy levels and a lack of parental support at home (the role of the home reading environment is also discussed further below).

> *We are an incredibly deprived school. Our [P]upil [P]remium is heading towards 70 per cent. Our reading data is [low] across year groups in the school.* (School coordinator)

Pupil engagement and responsiveness

Overall, pupil engagement was felt to be high. However, teachers found that pupils who were generally less engaged in school were less engaged with Reciprocal Reading. There were mixed views around whether pupils who contributed less to discussions were still benefiting from the programme. On the one hand, teachers felt that these pupils were still benefiting from being exposed to discussion. Others felt that this limited what their pupils got out of the programme:

> *We've recently had a Year 5 assessment week, and the children who were in that group, all bar one, have improved, some of them dramatically, and the one that hasn't improved is one of those boys that doesn't really engage in the sessions, even though I know he's so capable.* (Session Lead)

The fact that pupils found the summarising strategy the most difficult and hard to grasp was a recurring theme across the case studies. This was noted by researchers during the observations, expressed by teachers and reflected on by pupils who took part in a focus group, who said they sometimes forget what had happened. This may have been a barrier to achieving a higher positive effect on reading comprehension (estimated to be zero months).

Familiarity with the strategies may have influenced responsiveness. Teachers whose pupils were already using the language/strategies said that this had helped.

> *All of the language involved in the Reciprocal Reading is language that they already know and use.* (School coordinator)

Intervention as replacement

The retrospective survey suggests that, for many pupils participating in the intervention, Reciprocal Reading took place during usual reading sessions/other lessons. Around 64% of respondents indicated that pupils needed to be taken out of usual reading sessions or lesson time to attend some or all Reciprocal Reading sessions. Teachers told us that this was necessary due to timetable constraints. For those taken out of usual reading sessions Reciprocal Reading was not extra reading time, which may have had unintended consequences. However, teachers also said that this may have reduced the risk of backfire effects.

> *Because it has always been that session for us that they expect to be taken with a teacher, there have not been any of the self-esteem issues that you sometimes see historically with interventions.* (School coordinator)

Home reading environment

Evidence from the IPE supports the idea that family-level attitudes to reading and the home reading environment can influence the outcomes of Reciprocal Reading. School staff linked economic disadvantage to the quality of the home reading environment and prior reading skills, which could act as barriers to children making progress (this is noteworthy given that the average proportion of FSM pupils in the sample was 39.2%, higher than the national average of 24.6%, and may therefore, explain a barrier to effectiveness).

Cost

The cost evaluation aimed to estimate the total cost for schools to implement Reciprocal Reading. The average cost for one school to implement Reciprocal Reading for three consecutive years is £2,759 or approximately £77 per pupil per year, assuming the intervention is delivered to 12 pupils each year (total of 36 pupils).[36] The total cost is slightly higher than the estimated cost of the targeted intervention calculated in the efficacy trial (£2,436). The cost per pupil is higher as the previous evaluation assumed the intervention is delivered to six pupils per year.

Table 31: Financial cost of delivering Reciprocal Reading for schools

Type of cost	Item	Start-up or recurring?	Cost Year 1 Mean (Min–Max)	Cost Year 2 (forecasted mean)	Cost Year 3 (forecasted mean)	Total cost over three years	Total cost per pupil per year
Programme costs	Programme fee	Start-up	£1,965	0	0	£1,965	£54.58
Training costs	Training travel expenses	Start-up	£70 (0–£410)	0	0	£70	£1.94
Facilities, equipment, and materials	Materials	Recurring	£4 (0–£55)	£4.10	£4.3	£12.40	£0.34
	Photocopying / printing	Recurring	£9 (0–£29)	£9.30	£9.6	£27.90	£0.78
Personnel	Staff hire costs for delivery	Recurring	£160 (0–£1,389)	£169.12	£178.76	£507.88	£14.11
	External cover costs for training	Start-up	£175 (0–£560)	0	0	£175	£4.86
Total			£2,383	£183	£193	£2,759	£76.64

The estimated cost is the average of all financial expenditures related to the programme that were reported by the 15 schools in the sample. Table 31 presents the sample averages for the different types of financial costs.[37] The financial costs associated with the intervention include:

- **The programme fee.** Schools pay an initial fee between £1,000 and £1,500 (depending on form entry). FFT have a subscription model, which includes a discount if schools opt to sign-up for two or three years in total. Table 31 presents the total cost for a two-form entry school to sign-up to the programme for three years (£1,965, which includes the discount). All schools pay an annual fee of £300 to continue to access the programme resources and training. Note that for the trial, schools paid a reduced fee of £215.

- **Training costs.** Most schools in the trial attended in person training (two days), which required travel to a venue. Access to training is included in the fee but schools had to pay for any travel costs. The average cost for travel is shown in Table 31. This is classed as a one-off start-up cost as new staff tend to be trained up online.

- **Equipment/materials.** Schools receive all the materials including anthologies as part of their subscription. Schools can rotate these. They may purchase additional texts, but this is not essential. Costs of printing/photocopying online resources were minimal and are shown in Table 31.

[36] This is the mean average number suggested by FFT (see Table 5). In the trial, the mean was 14 pupils per school.
[37] Some of these averages mask large differences across schools and we report the range of values for each component to reflect this.

- **Personnel.** Some schools who took part in the trial incurred financial costs where they paid for external cover/additional hours to allow for training/delivery.

The cost forecasts for years two and three include inflation adjustments,[38] assuming inflation remains constant in the next two years. Economic and market fluctuations could affect our total cost estimates primarily through teacher wage increases, as other financial costs in the programme are minimal. Our estimates assume schools pay the programme fee—the other main expense—upfront in year one to cover all three years. Since we expect these economic factors to have limited impact, no additional analysis on future cost uncertainty is necessary.

Table 32: Cumulative costs of Reciprocal Reading (assuming delivery over three years)

Programme	Year 1	Year 2	Year 3
Reciprocal Reading	£2,383	£2,566	£2,759

As well as financial costs, we collected data on time spent by staff on the programme. Staff had to spend time on the following activities:

- training—two training days;
- trial management tasks;
- preparation/planning;
- delivery; and
- internal cover for training/delivery.

We report the total time spent on tasks by school coordinators (usually members of the SLT) and session leads (teachers/teaching assistants). In most schools in the sample there were two session leads (each delivering one group each), however, in four schools there was only one member of staff delivering sessions. In most schools, the coordinator tasks were performed by one person, but this was sometimes shared out. Table 33 shows the time spent by staff on setting up the programme, including attending training sessions/supporting meetings, providing cover for staff who attended training, and selecting pupils to take part. These are considered to be one-off costs.

Table 33: Staff time spent on set-up (total)

		School coordinator	Session leads
		Mean number of hours in total (Min–Max)	Mean number of hours in total (Min–Max)
Training	Attending training days (including travelling to venue)	16.5 (8.5–28)	27.4 (14–40)
	Attending programme briefing session	0.9 (0–1)	1.1 (0–2)
	Attending online support meetings	1.2 (0–2)	2.0 (0–4)

[38] According to the ONS, Consumer Price Index inflation was 3.3% in 2024, and the public wage annual growth rate was 4.7%. We assume these figures stay constant in years two and three. Sources: www.ons.gov.uk/economy/inflationandpriceindices/timeseries/l55o/mm23 and www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/averageweeklyearningsingreatbritain/february2025

	Accessing ad hoc support	0.4 (0–3)	0.2 (0–2)
Teacher cover for training	Time spent by other school staff covering Reciprocal Reading team while they were attending training	1.8 (0–20)	7.3 (0–40)
Trial management	Screening and selecting pupils to take part	1.0 (0–2.5)	0.5 (0–5)
Total		21.8 (11–45)	38.1 (15.8–86)

Table 34 shows time spent by staff delivering the intervention on a weekly basis. This is based on delivering to two groups of six pupils. This only covers tasks for which estimated times were provided per week (rather than over the full delivery time).

Table 34: Staff time spent on implementation (recurring, per week)

		School coordinator Mean amount of time weekly (Min–Max)	Session leads Mean amount of time weekly (Min–Max)
Delivery	Planning e.g. reviewing example session plans, planning sessions independently, preparing materials, reading texts	0.07 hours (0–1)	0.9 hours (0–2)
	Delivery—facilitating the sessions	0 hours	2.4 hours (0–5)
Trial management	Recording pupil attendance and submitting figures	0.1 hours (0–1)	0.3 hours (0–0.5)

Table 35 shows a breakdown of staff time spent on implementation in total, assuming a 12-week intervention (minimum dosage recommended by FFT). For 12 weeks of delivery, the average time commitment is approximately five hours for a school coordinator, and 45 hours for session leads (this could be split between two members of staff). For session leads, most of the time is used for planning and delivery. For school coordinators, time is used for a combination of scheduling/timetabling, internal meetings, and monitoring attendance.

Table 35: Staff time spent on implementation (total)

		Year 1	
		School coordinator Total mean number of hours (Min–Max total)	Session leads Total mean number of hours (Min–Max total)
Delivery	Scheduling / timetabling sessions, e.g. timetable amendments	0.9 (0–5)	0.1 (0–1)
	Staffing arrangements, e.g. finding staff cover (if needed)	0.4 (0–1)	0
	Internal check-ins – meetings among staff involved in delivery	0.9 (0–2)	1.1 (0–4)
	Observations – observing sessions led by other staff, recording observations / writing up notes	0.6 (0–2)	0.07 (0–1)
	Planning, e.g. reviewing example session plans, planning sessions independently, preparing materials, reading texts	0.8[a] (0–12)	11.4[a] (0–24)
	Delivery – facilitating the sessions	0[a]	28.4[a]

			(0–60)
Trial management	Recording pupil attendance and submitting figures	1.5[a] (0–12)	3.7[a] (0–6)
Internal cover for delivery	Time spent by other school staff covering the Reciprocal Reading team while they were delivering	0	0.07 (0–1)
Total		5.2 (1–18.8)	44.9 (17.4–74)

[a] *These figures were collected as a per week figure. Totals assume 12 weeks of delivery.*

Conclusion

Table 36: Key conclusions

Key conclusions	
1.	Pupils in Reciprocal Reading schools made one month's more progress in reading on average, compared to pupils in other schools. This result has a high security rating.
2.	Among pupils eligible for free school meals (FSM), those in Reciprocal Reading schools made no additional month's progress in reading, on average, compared to pupils in other schools. These results have a lower security than the overall findings because of the smaller number of pupils.
3.	Supplementary analysis revealed that intensity matters. Pupils who received both the recommended intensity and minimum number of sessions (at least 20 sessions over 12 weeks or less, as opposed to more spaced-out delivery) received the equivalent to two months' progress, as compared to the average pupil in the intervention group.
4.	Teachers perceived that training was of a high quality and followed the delivery model closely. High attendance and positive engagement with the training sessions meant that teachers delivered the intervention as intended.
5.	Teacher observations were of a high quality, with teacher attitudes towards the intervention reported as overwhelmingly positive, and teachers reporting high levels of pupil engagement given the interactive nature of the intervention.

Impact evaluation and IPE integration

Evidence to support the logic model

There is strong evidence to support the activities in the logic models for the training programme and the programme for pupils. The fidelity data collected suggests that trainers followed the training plan closely. There were only minor differences between sessions, which mainly related to the order of topics or, more exceptionally, the level of detail covered in certain sections. Fidelity data also suggests that teachers adhered closely to the model when delivering the programme with their pupils.

There is also strong evidence in support of the hypothesised mechanisms. Analysis of assessments with teachers before and after training shows large increases in their self-reported confidence in implementing Reciprocal Reading, self-reported knowledge of Reciprocal Reading concepts, and actual increases in knowledge. Interviews with teachers further support these findings. Most teacher survey respondents (82%) believed that, compared to their school's usual teaching, Reciprocal Reading was more effective at boosting reading skills. These survey respondents said the most important factor driving this improvement was the fact that the intervention gave pupils a more structured approach to reading through the strategies. This supports the theory that teaching children the four strategies helps them to better comprehend what they are reading. There was also evidence from interviews that the strategies had given pupils a framework to apply to texts more widely (outside of the sessions). In some cases, teachers said their pupils had used the strategies and approach (breaking the text up into manageable sections in order to discuss and monitor understanding) during practice SAT papers (though the impact analysis does not find a positive effect for Year 6 pupils on Key Stage 2 SAT reading scores).

Pupil engagement with the intervention was high. Around 98% of retrospective survey respondents said that pupils had been 'very engaged'/'somewhat engaged'. This was also reflected in the case studies. Overall, pupils were highly engaged in the sessions that were observed, albeit with some differences between pupils.

There was moderate evidence from teacher interviews to support five of the hypothesised moderators. Teachers believed that the following factors could have affected the effectiveness of the programme: i) having EAL; ii) the pupil's prior reading skills; iii) whether the pupil was eligible for FSM; iv) the pupil's engagement in sessions; and v) the home reading environment. It is, however, hard for teachers to know these things with any certainty, purely from their observations. It should also be noted that the subgroup analysis does not really illuminate this discussion. It suggests that pupils eligible for FSM either received no effect or a similar effect to all pupils (depending on the type of analysis), but there is a fairly high level of uncertainty around all estimates that probably explains the discrepancy in results.

The evidence relating to the primary outcome suggests that pupils in the intervention received a small positive effect on their overall reading attainment, relative to their peers in the control group (noting that the results are consistent with a range from no effect to a slightly larger positive effect). When we break this down into the subscales of the NGRT, we see that the overall effect may have been driven by small improvements in reading accuracy and basic comprehension, as measured by the sentence completion score. We find no effect on passage comprehension. This is an important finding because passage comprehension is considered by the developers to be the closest measure of the programme's main target outcome (the rationale for including it as a dual primary outcome in the efficacy trial). So, the intervention does seem to be improving pupils' reading attainment but may not be producing the exact intended outcomes. This distinction should not be overstated however, as the confidence is quite wide on the passage comprehension estimate. It is also worth noting that the IPE analysis found that pupils and parents both believed that the intervention had a positive effect on comprehension. Teachers also saw improvements in pupil comprehension but were more cautious about attributing this to the programme.

Interpretation

This study provides further evidence that the Reciprocal Reading programme has positive effects on pupils' overall reading attainment, as well as their reading accuracy and basic comprehension (estimated to be one month of progress for both outcomes, noting that these results are consistent with a range from no effect to slightly larger positive effects). We cannot say what kind of effect, if any, the programme has on passage comprehension or Key Stage 2 SAT reading scores due to the large amount of uncertainty in the results. Where a positive impact was detected, it seems to have been small on average; approximately half the size of the impact found in the previous efficacy trial (O'Hare *et al.,* 2019). We can be confident that this comparison is a valid one. The present (effectiveness) trial used the same outcome measures, the same experimental design, and collected detailed data on the five components of fidelity—adherence, dosage, quality of delivery, participant responsiveness, and programme differentiation—to explore in detail what was delivered.

The observed drop-off in impact between the efficacy and the effectiveness trial is to be expected. It is common for programmes to be less effective on average when delivered at scale (List, 2022). This can be due (in general) to greater variation in the schools, teachers, and pupils taking part, as well as the difficulty in maintaining quality of delivery at scale (Maxwell *et al.,* 2021; Hall *et al.,* 2025). In the case of this trial, we found evidence of two factors being particularly influential.

First, when we re-estimate the impact for pupils who received a minimum defined dosage, we find that intensity matters. Pupils who received both the recommended intensity and minimum number of sessions (at least 20 sessions over 12 weeks or less) received a larger effect, equivalent to two months progress, as compared to the average pupil in the intervention group. On average, pupils in the efficacy trial received nearly double the number of sessions than pupils in this trial (52 sessions over 26 weeks vs 28 sessions over 15 weeks) so this could go some way to explaining the difference in estimated impacts between the two trials. Very few schools delivered near to 52 sessions in this present trial. This could be an indication that this number is hard to deliver by the wider range of schools included but equally could be a result of clearer guidelines on dosage from FFT. Either way, a more efficient route to higher impact could be through delivering the recommended intensity of sessions. Our compliance analysis shows that pupils who received both the recommended intensity and minimum number of sessions (at least 20 sessions over 12 weeks or less) received a larger effect, equivalent to two months progress (equal to the effect seen in the efficacy trial).

Second, a lot of targeted reading support of varying kinds was clearly delivered to pupils in the control group. A total of 38 schools in the control group used reading interventions from other organisations including PiXL, FFT (Tutoring with the Lightning Squad), and Reading Plus. This may also help to explain the smaller impact found in this trial. However, while the efficacy trial captured information on the school comprehension ethos and reading strategy behaviours in control schools, it did not collect details on the types of targeted support/interventions that were delivered, which limits our ability to make comparisons around the level of targeted support control pupils in the two trials received.

We can say with fairly high confidence that the programme operated as intended—both in terms of the training programme for teachers and the reading programme for pupils. We can be confident, therefore, that the estimated effects are coming from the programme as it has been designed.[39]

Limitations and lessons learned

This study was a well-executed randomised design, with an MDES and level of attrition that meet the EEF's standards for high validity. Covariate data was also collected with very high levels of completeness. The outcomes are all measured with established tools that have strong validity and reliability. In the IPE, good evidence has been provided to suggest that the intervention was implemented with high fidelity to the programme structure and resources. Despite this strong design and execution, we have explored three main potential sources of bias in the analysis:

1. **Month of test.** This is the month in which the pupil took the NGRT endline test, taking the values of 1 to 12 to correspond with the calendar months. It was not a baseline covariate, but we checked the balance of this variable to assess the risk of bias. The average month of testing was 5.12 in the intervention group, meaning that pupils in the group took the test, on average, just after the beginning of May. This figure in the control group was 4.95, meaning that pupils in the group took the test, on average, at the end of April. This mean difference of 0.17 months is equal to about five days. We do not expect this small difference in the average endline test date to introduce substantial bias but, as pre-specified, month of test was included as a covariate in the analysis to account for the spread of dates over which pupils took the test.
2. **Adaptive testing.** There is a theoretical risk of bias in the passage comprehension score, as pupils need to show a basic level of reading to access the passage comprehension part of the test (see page 11 of the Statistical Analysis Plan for more information; Torres Blas and Taylor, 2019). With only 1.1% of pupils with an NGRT endline score missing a passage comprehension score, and no significant difference in the probability of having a score, we conclude that there is little to no risk of bias in the passage comprehension results.
3. **Model specification.** As pre-specified, all primary outcome and secondary outcome models were re-estimated without covariates. Additionally, the primary outcome was re-estimated with a model without pupil-level covariates, including only treatment assignment, baseline attainment, and trial design characteristics (the randomisation batch). This model allows for more comparability with other trials by the EEF. Overall, the results of the main analysis are robust to these changes in model specifications.

The NGRT outcomes are short-term measures of change. This is a strength because it allows us to evaluate any change shortly after the programme has ended and before any washout may occur. But it is also a weakness because we are unable to comment on the longer-term effects of the programme.

To assess the generalisability of the findings from this trial, we compare the characteristics of our sample to the national averages. This shows that the sample of schools in the trial was more urban, rated better by Ofsted, and had a more economically deprived intake of pupils, compared to all English primary schools. The location, school performance, and intake could moderate the effects of the intervention. The effects identified in our analysis may therefore, be different if the treatment were scaled up to all schools in England. The IPE findings suggest that the quality of the trainers, the teachers, and teaching assistants who deliver the programme are important factors in the success of the programme. Recruiting high-quality staff is probably easier in urban areas where the population density is higher. So, if the programme were scaled-up nationally, an increase in the proportion of rural schools in the sample may lead to a reduction in the effects.

The IPE findings were limited by the quality of data that we managed to collect from parents. While findings from the interviews with parents helped to contextualise some of the findings from interviews with teaching staff about perceived

[39] A key difference between the efficacy and effectiveness trial was that a train-the-trainer model was developed and implemented. While this took place prior to this trial and is therefore, outside of the scope of the evaluation, the high fidelity observed suggests that the shift to a wider pool of trainers has not had a significant impact.

outcomes, the data was more limited due to parents' low awareness of the programme and challenges recognising changes in their child's reading habits and abilities. They therefore, feature less prominently in the findings than other methods.

Future research and publications

We suggest that future research in this field should focus on how to increase the effectiveness of Reciprocal Reading at scale. The larger effects estimated in previous studies, suggest that the programme can be more effective. The IPE suggests that this can be achieved with a greater focus on ensuring that the recommended dosage and intensity are delivered, while maintaining the high levels of adherence and quality. A scale-up evaluation of the programme could support this.

References

Cappellini, C., Torres Blas, N. and Taylor, P. (2022) *'Reciprocal Reading Effectiveness Trial: Evaluation Protocol'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Shared-with-EEF-TP-Reciprocal-Reading-V5_CLEAN-VERSION.pdf?v=1752510135 (accessed 04 February 2026).

Clark, C. and de Zoyah, S. (2011) 'Mapping the Interrelationships of Reading Enjoyment, Behaviour and Attainment'. London: National Literacy Trust. Available at: https://literacytrust.org.uk/research-services/research-reports/mapping-interrelationships-reading-enjoyment-attitudes-behaviour-and-attainment-2011/ (accessed 04 February 2026).

Cockerill, M., Thurston, A., O'Keeffe, J. and Taylor, A. (2021) 'A Phase 3 Definitive RCT of Reciprocal Reading in High Schools in England'. *International Journal of Educational Research,* 109, 101854. https://doi.org/10.1016/j.ijer.2021.101854

Cockerill, M., O'Keeffe, J., Thurston, A. and Taylor, A. (2022) 'Reciprocal Reading for Struggling Readers: An Exemplar of Evidence Implementation in Schools'. *Review of Education,* 10: 1, e3332. https://doi.org/10.1002/rev3.3332

Cockerill, M., Thurston, A., O'Keeffe, J. and Chiang, T. H. (2025) 'Results From a Phase 3 Definitive Trial of Reciprocal Reading in English High Schools. *International Journal of Educational Research*, 130, 102546. https://doi.org/10.1016/j.ijer.2025.102546

Crawford, C. and Skipp, A. (2014) *'LIT Programme Evaluation Report and Executive Summary October 2014'*. London: Educational Endowment Foundation.

Data Protection Act 2018, c.12. Available at: www.legislation.gov.uk/ukpga/2018/12/contents (accessed 04 February 2026).

Department for Education (DfE). (2022) *'Academic Year 2021/22: Key Stage 1 and Phonics Screening Check Attainment'*. GOV.UK. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-1-and-phonics-screening-check-attainment/2021-22 (accessed 15 July 2025).

Education Endowment Foundation (EEF). (2023) *'Cost evaluation guidance for EEF evaluations'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/Cost-Evaluation-Guidance-Feb_2023.pdf?v=1770203122 (accessed 21 July 2025).

Fetters, M.D., Curry, L.A. and Creswell, J.W. (2013) 'Achieving Integration in Mixed Methods Designs—Principles and Practices'. *Health Services Research,* 48: (6pt2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

General Data Protection Regulation (GDPR). (2016) *'Council Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (United Kingdom General Data Protection Regulation) (Text with EEA relevance)'*. Available at: www.legislation.gov.uk/eur/2016/679

GL Assessment. (2025a) *'Using NGRT in Primary Schools'*. Brentford: GL Assessment. Available at: www.gl-assessment.co.uk/products/new-group-reading-test/ (accessed 21 July 2025).

GL Assessment. (2025b) *'Technical Guidance'*. Brentford: GL Assessment. Available at: https://support.gl-assessment.co.uk/knowledge-base/assessments/ngrt-support/general-information/technical-guidance (accessed 15 July 2025).

Gough, P.B. and Tunmer, W.E. (1986) 'Decoding, Reading, and Reading Disability'. *Remedial and Special Education,* 7: 1, 6–10. https://doi.org/10.1177/074193258600700104

Hall, A., Baan, A.-M., Taylor, P. and Lewis, J. (2025) 'EEF Scaling Framework'. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/EEF_scaling-framework_v.1.1.0.pdf?v=1752497134 (accessed 14 July 2025).

Hoffmann, T.C., Glasziou, P.P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D., Macdonald, H., Johnston, M., Lamb, S.E., Dixon-Woods, M., McCulloch, P., Wyatt, J.C., Chan, A-W. and Michie, S. (2014) 'Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide'. *BMJ,* 348. https://doi.org/10.1136/bmj.g1687

Horowitz, J.L. and Manski, C.F. (1998) 'Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations'. *Journal of Econometrics,* 84: 1, 37–58. https://doi.org/10.1016/S0304-4076(97)00077-8

Lee, D.S. (2009) 'Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects'. *The Review of Economic Studies,* 76: 3, 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

List, J.A. (2022) *'The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale'.* New York, NY: Crown Currency.

Maxwell, B., Stiell, B., Stevens, A., Demack, S., Coldwell, M., Wolstenholme, C., Reaney-Wood, S. and Lortie-Forgues, H. (2021) *'Review: Scale-Up of EEF Efficacy Trials to Effectiveness Trials'.* London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/scale-up-of-eef-efficacy-trials-to-effectiveness-trials (accessed 14 July 2025).

Montgomery, J.M., Nyhan, B. and Torres, M. (2018) 'How Conditioning on Post-Treatment Variables Can Ruin Your Experiment and What to Do About It'. *American Journal of Political Science,* 62: 3, 760–775. https://doi.org/10.1111/ajps.12357

O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., Biggart, A., Connolly, P. and Bower, C. (2019) *'Reciprocal Reading: Evaluation Report'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/reciprocal-reading (accessed 14 July 2025).

Ritchie, J., Lewis, J., McNaughton Nicholls, C. and Ormston, R. (Eds.). (2013). Qualitative research practice: A guide for social science students and researchers. London, Thousand Oaks, CA: Sage Publications Ltd.

Rosenshine, B. and Meister, C. (1994) 'Reciprocal Teaching: A Review of the Research'. *Review of Educational Research,* 64: 4, 479–530. https://doi.org/10.3102/00346543064004479

Thurston, A., Cockerill, M., Chiang, T.-H., Taylor, A. and O'Keeffe, J. (2020) 'An Efficacy Randomized Controlled Trial of Reciprocal Reading in Secondary Schools'. *International Journal of Educational Research*, 104, 101626. https://doi.org/10.1016/j.ijer.2020.101626

Torres Blas, N. and Taylor, P. (2019) *'Reciprocal Reading Effectiveness Trial: Statistical Analysis Plan'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/reciprocal-reading_statistical-analysis-plan_v.1.1.0.pdf?v=1754942543 (accessed 04 February 2026).

Appendix A: EEF cost rating

Appendix A Table 1: Cost rating

Cost rating	Description
£ £ £ £ £	Very low: less than £80 per pupil per year.
£ £ £ £ £	Low: up to about £200 per pupil per year.
£ £ £ £ £	Moderate: up to about £700 per pupil per year.
£ £ £ £ £	High: up to £1,200 per pupil per year.
£ £ £ £ £	Very high: over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: New Group Reading Test (NGRT) by GL Assessment

Please use this template to assign a separate security rating for each primary outcome. Secondary outcome analysis and/or subgroup analyses are NOT included in the security ratings unless otherwise stated.

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	**MDES**	**Attrition**					5
5	Randomised design	<= 0.2	0-10%					
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 – 0.29	11– 20%					
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 – 0.39	21– 30%			Adjustment for threats to internal validity **[-1]**		4
2	Design for comparison that considers selection only on some relevant confounders	0.40 – 0.49	31–40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 – 0.59	41–50%					
0	No comparator	>=0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	Balance is excellent at randomisation. As analysed remains very well balanced, with the exception of month, which will have little overall impact.
Threat 2: Concurrent interventions	Moderate	The usual practice survey identifies use of other interventions in intervention schools and additional targeted support in control schools. While there is no evidence that treatment allocation

		was correlated with differential uptake of these interventions, it has also not been possible to analyse and rule out partial and untracked delivery of Reciprocal Reading, which may have masked effects. Likely direction—diluted ability to detect effect.
Threat 3: Experimental effects	Moderate	There is evidence of minor contamination with 14.6% of control schools (who returned retrospective survey) stating full or partial implementation of Reciprocal Reading. Unclear whether compensatory or part of usual practice of staff previously trained. Likely direction—diluted ability to detect effect.
Threat 4: Implementation fidelity	Low	The intervention was delivered with fidelity. There were some minor adaptations.
Threat 5: Missing data	Low	Total missing data is very low.
Threat 6: Measurement of outcomes	Low	Tests are well known. There was a small issue around reading comprehension scores not being available for those who scored very low, but this was handled appropriately in the design. There was an unreported number administered by school staff at baseline or endline, but again this appears to be minor. Some retakes (20%) were necessary due to IT issues, etc., but again this was handled appropriately. There doesn't appear to be much in the way of floor or ceiling effects.
Threat 7: Selective reporting	Low	No evidence of selective reporting.

- **Initial padlock score.** 5 Padlocks – the study was powered to detect a small effect (MDES of 0.12 on the overall sample) and attrition in each trial arm is under 10%.
- **Reason for adjustment for threats to validity.** Three threats are classified as moderate risks, with two of them with the same likely direction of bias.
- **Final padlock score.** Initial score adjusted for threats to validity = 4 Padlocks.

Appendix C: Changes since the previous evaluation

Appendix C Table 1: Changes since the previous evaluation

	Feature	Efficacy to effectiveness stage
Intervention	Intervention content	No change.
	Delivery model	To scale delivery from the efficacy trial to the effectiveness trial, FFT trained a new cohort of trainers, who then trained the teachers and teaching assistants who delivered the programme to pupils. In the efficacy trial, the training for teachers and teaching assistants was delivered by a small team of highly experienced trainers who now act as FFT senior trainers; responsible for training trainers.
	Intervention duration	On average, pupils in the efficacy trial received nearly double the number of sessions than pupils in this effectiveness trial (52 sessions over 26 weeks vs 28 sessions over 15 weeks). This was in part down to a wider window for endline testing, due to the large number of schools in the effectiveness trial. Both trials exceeded the minimum dosage specified by the developers (24 sessions over 12 weeks).
Evaluation	Eligibility criteria	No change.
	Level of randomisation	No change.
	Outcomes and baseline	The same outcomes were collected in both trials. The efficacy trial had two primary outcomes. This effectiveness trial classified one of those outcomes (passage comprehension) as a secondary outcome.
	Control condition	A lot of targeted reading support of varying kinds was clearly delivered to pupils in the control group in the current trial. Around 38 schools in the control group used reading interventions from other organisations including PiXL, FFT (Tutoring with the Lightning Squad), and Reading Plus. This may also help to explain the smaller impact found in this trial. However, while the efficacy trial captured information on the school comprehension ethos and reading strategy behaviours in control schools, it did not collect details on the types of targeted support/interventions that were delivered, which limits our ability to make comparisons around the level of targeted support control pupils in the two trials received.

Appendix D: Effect size estimation

Appendix D Table 1: Effect size estimation

Outcome	Unadjusted differences in means	Adjusted differences in means	Intervention group		Control group		Pooled variance	Population variance (if applicable)
			n (missing)	Variance of outcome	n (missing)	Variance of outcome		
NGRT overall reading score	2.865	2.504	1,907 (211)	1881.106	1,971 (174)	2255.956	2071.626	N/A
NGRT passage comprehension score	2.059	1.287	1,869 (249)	2117.985	1,924 (221)	2333.438	2227.274	N/A
NGRT sentence completion score	2.500	2.450	1,907 (211)	2066.152	1,971 (174)	2374.678	2222.962	N/A
Key Stage 2 SAT reading score	-0.370	-0.308	1,015 (1,103)	41.900	1,035 (1,110)	38.112	39.988	N/A

N/A = not applicable.

Further appendices:

See separate attachment on the EEF website that covers the following.

Appendix E: FFT School – Memorandum of Understanding (MOU)
Appendix F: Randomisation code
Appendix G: Parent information sheet and withdrawal form – trial
Appendix H: Parent information sheet and withdrawal form – focus group
Appendix I: BIT privacy notice
Appendix J: Pre- and post-training analysis – knowledge quiz results
Appendix K: Deviations from the protocol

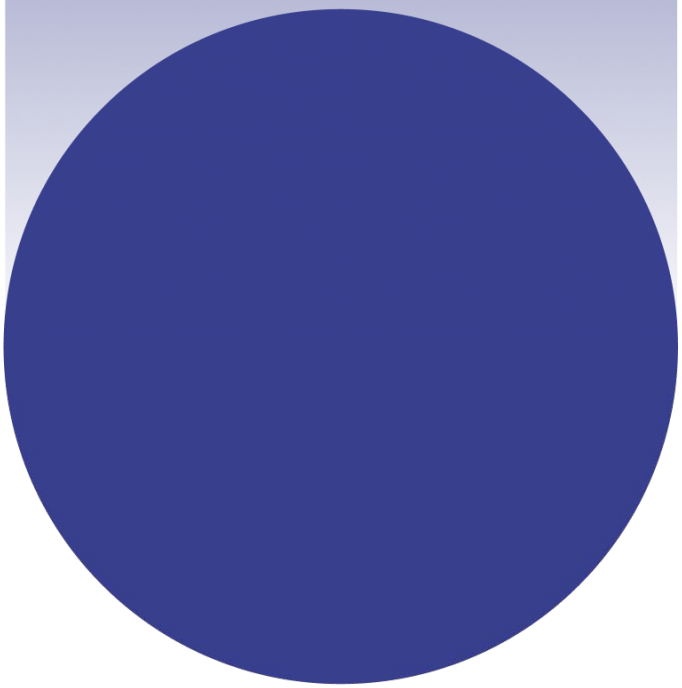
.

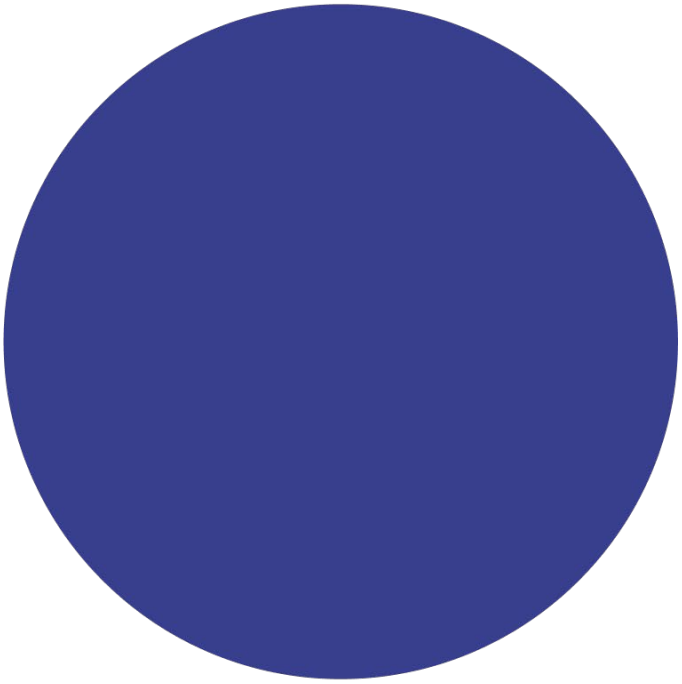
You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit https://nationalarchives.gov.uk/doc/open-government-licence/version/3 or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at https://educationendowmentfoundation.org.uk





Education Endowment Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn