Comparing the effectiveness of two reciprocal reading comprehension interventions for primary school pupils in disadvantaged schools

Received: 9 March 2022 | Accepted: 8 June 2023

DOI: 10.1111/bjep.12623

ARTICLE

Comparing the effectiveness of two reciprocal reading comprehension interventions for primary school pupils in disadvantaged schools

Liam O'Hare[1] | **Patrick Stark**[1] | **Maria Cockerill**[1] | **Katrina Lloyd**[1] | **Sheila McConnellogue**[1] | **Aideen Gildea**[1] | **Andy Biggart**[1] | **Christine Bower**[1] | **Paul Connolly**[2]

[1]Queen's University Belfast, Belfast, UK

[2]Ulster University, Coleraine, UK

Correspondence
Liam O'Hare, Queen's University Belfast, 69-71 University St, Belfast, UK.
Email: l.ohare@qub.ac.uk

Funding information
Education Endowment Foundation

Abstract

Background: Effective reading comprehension teaching is an aspiration of education systems across the world. Teaching incorporating reciprocal reading theory and evidence is an internationally popular approach for improving comprehension.

Aims: This paper uses two large cluster randomized controlled trials of similar reciprocal reading interventions implemented in different ways to compare their effectiveness.

Sample: The two interventions had the same teacher professional development, reciprocal reading activities and dosage/exposure, but varied in their implementation, with one delivered as a whole-class ('universal') version for pupils aged 8–9 years and the other a small group ('targeted') version for pupils aged 9–11 years with specific comprehension difficulties.

Methods: Two large-scale cluster RCTs were conducted in 98 schools with $N = 3699$ pupils in the universal trial and $N = 1523$ in the targeted trial.

Results: Multi-level models showed significant effects for the targeted version of the intervention on pupil reading comprehension ($g = .18$) and overall reading ($g = .14$). No significant effects were found for the whole class version. A sub-group analyses of disadvantaged pupils showed the targeted intervention's effects were even larger on reading comprehension ($g = .25$).

Conclusions: The evidence suggested that this reciprocal reading intervention worked best when implemented in

small groups and targeted for pupils with specific comprehension difficulties and particularly for pupils in disadvantaged circumstances.
Comments: This evaluation shows that even if a reading comprehension intervention is underpinned by strong theory and evidence-based practice, its effectiveness can still depend on implementation choices.

KEYWORDS
comprehension, implementation, literacy, randomized controlled trial, reciprocal reading

BACKGROUND

Policymakers in the United States and United Kingdom have placed considerable emphasis on reducing those leaving elementary-aged schooling with poor levels of literacy. In the United Kingdom, the pedagogical focus has been on systematic synthetic phonics, which has a strong evidence base (Ehri et al., 2001) but there is increasing recognition that some children successfully learn to decode text but struggle with text comprehension. This was recognized by the National Reading Panel (2000) in the United States and as a result they recommended reading comprehension strategies as one of the five key components of reading strategy instruction.

As a result of these developments, over the last few decades, the teaching of reading strategies is increasingly a standard part of international reading curricula in elementary and secondary education (Department for Education, 2013a; Okkinga et al., 2018; Pressley, 2002). An example of a reading comprehension strategy intervention is reciprocal teaching. Reciprocal teaching is now essentially a worldwide approach (Palincsar, 2013). However, reciprocal teaching type interventions have not been widely implemented in the United Kingdom as a discrete intervention until recently, but it is now recommended by the UK's Education Endowment Foundation (EEF) teaching toolkit (Education Endowment Foundation, 2022a).

Reciprocal teaching is informed by research and development in the United States that was led originally by Palincsar and Brown (1984). Although there are a range of interventions based on the principles of reciprocal teaching, the core factors are comprehension strategy instruction through modelling of strategies by an adult and student practice of the modelled strategies (De Corte et al., 2001; Palincsar & Brown, 1984). Often, learners are explicitly taught four sequenced evidence-based strategies, which effective readers are known to use to help them derive meaning from texts (Palincsar & Brown, 1984). These are: (1) predicting what will happen next; (2) clarifying new or unfamiliar words; (3) asking questions about the text and (4) summarizing what has been read. The modelling is interactive in that, whilst adult models these strategic actions with the text, there is also a collaborative problem-solving approach to the text (Palincsar & Brown, 1984). Predicting is used first in the present study, which differs from Palincsar and Brown where summarizing is used first. The rationale for this is to develop their inference practice using a few clues from the text provided (e.g. text title, the first few sentences and building on text knowledge from a previous session). There is substantive evidence that the ability to draw inferences predetermines reading skills: that is, poor inferencing causes poor comprehension and not vice versa (Kispal, 2008). FFT Literacy also suggests using prediction first promotes oral discussion from the outset in a group of students who would not be used to this kind of cooperative working together on the reading of a text. For these reasons, the FFT reciprocal reading strategy cycle evaluated in this study begins with predicting and ends with summarizing.

Previous meta-analyses have claimed positive effects for general reading strategy instruction (Chiu, 1998; Swanson, 1999) including comprehension strategies (Sencibaugh, 2007) and reciprocal

teaching (Rosenshine & Meister, 1994). In the latter meta-analysis of 16 reciprocal teaching studies an average effect size of .32 was found for reading and .88 for reading comprehension (Rosenshine & Meister, 1994). Despite the large international research base on reading comprehension strategy instruction, the EEF (2022b) has stated that to date there are few UK-based evaluations of comprehension programmes and those available generally show lower effects on pupil outcomes than the international (mainly United States) literature.

One of the possible explanations of difficulties in replicating large effect sizes of reading comprehension interventions is variation in implementation methods. There are some key reciprocal teaching implementation issues that are worthy of note in the literature. The evidence base for reciprocal teaching generally centres on implementation with specific targeted populations (students with literacy or learning difficulties) and in a recent meta-analysis of 52 studies, Okkinga et al. (2018) showed larger effect sizes for low achievers in reading, rather than typically developing students. Effectiveness of universal (whole-class) approaches has also been difficult to demonstrate (De Corte et al., 2001), with research suggesting the intervention is challenging to deliver in a whole-class setting and still maintains intervention quality (Okkinga et al., 2018). Also, some studies report that teachers find it hard to instruct students in strategic thinking (Duffy, 1993). In addition, Okkinga et al. (2018) report that teachers are generally less successful than researchers at delivering reading strategy instruction. This all suggests that reciprocal teaching requires a specialist understanding of reading comprehension itself and effective comprehension strategies. Therefore, additional specialist professional development for teachers may be required to deliver reciprocal teaching approaches effectively.

There is also a lack of clarity about the optimal age at which reciprocal teaching is beneficial. However, Okkinga et al. (2018) did report that reading strategy intervention is particularly beneficial for students aged 8–14 (even in a whole-class format). Regardless, age, developmental stage and cognitive capacity may be limiting factors for when students can receive a reading strategy-based intervention.

Another implementation issue is the level of disadvantage faced by reading programme participants. Pupil disadvantage has been shown to be a significant predictor of reading level as assessed in terms of accuracy, comprehension and rate (McPhillips & Sheehy, 2004). However, randomized controlled trials (RCTs) of language interventions in the United Kingdom have shown greater levels of reading improvement for pupils eligible for free school meals (FSM – a proxy for disadvantage) from those not eligible for FSM (Maxwell et al., 2014; Thurston et al., 2016).

Finally, study methodology has also had an impact on reciprocal teaching evaluation findings in the literature. Okkinga et al. (2018) meta-analyses reported smaller effect sizes (mean ES = .19) from studies using standardized tests compared to researcher-created proximal reading comprehension tests (mean ES = .43).

THE PROGRAMME

The Fisher Family Trust Literacy (FFT) *Reciprocal Reading* (RR) programme is based on the principals of reciprocal teaching. The RR programme is implemented with practicing Teachers/Teaching Assistants in mainstream United Kingdom settings and supports them in delivery of RR to pupils aged 8–11 years. In essence, it is a continuing professional development programme (CPD) for primary school educators. All Teachers/Teaching Assistants involved in delivering the programme receive 2 days' off-site CPD from RR instructors. The CPD covers the knowledge, skills and understanding that practitioners need to deliver RR. The training emphasizes the importance of all readers applying the four common reciprocal teaching strategies of predict, clarify, question, and summarize rather than individual readers taking on separate roles as in other reciprocal teaching programmes, for example, Reading Rockets (2023), where one student takes on the role of predictor, another of clarifier etc. The CPD considers the difficulties involved in comprehension and how RR addresses these, how to identify children who will benefit most from RR, as well as giving teachers opportunities to practice using the strategies. Additional sessions include considering how to build challenge into the

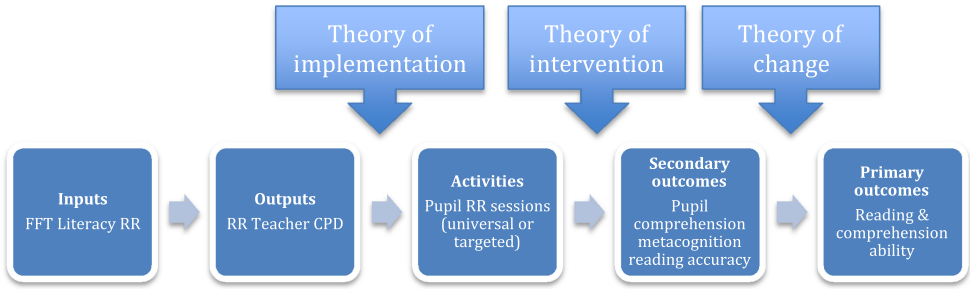

FIGURE 1 Simplified theory of change and logic model for FFT reciprocal reading programme.

lesson without losing the emphasis on children's talk and reviewing how best to enhance reciprocal teaching within a comprehension curriculum. FFT also provide in-situ support throughout delivery of RR to pupils. For example, the CPD is supplemented by school visits which can include demonstration, review of developing practice and advice on embedding the approach into classroom practice.

For this study teachers were instructed in the implementation of two different versions of the RR programme, namely: a whole-class ('universal') version and a 'targeted' programme. The universal version of the RR programme was for elementary school-aged pupils in England (8- to 9-year-old children equivalent to UK Year 4 and US Grade 3). This version was to be delivered to a whole class for 20 min at least once per week for a minimum of 12 weeks over a school year (i.e. a minimum of 240 min). The 'targeted' programme was for pupils aged 9–11 years old (UK Years 5 and 6, and US Grade 4 and 5) who were identified as having poor reading comprehension but good decoding skills. This version was to be delivered to groups of approximately six pre-selected pupils with comprehension difficulties, through at least two sessions of 20 min per week for 12 weeks (minimum recommended 480 min). Both versions of the RR programme were delivered by teachers and teaching assistants who had received the RR CPD training.

The programme's logic model (summarized in Figure 1) is a graphical representation of the programme's inputs, outputs, activities and outcomes. Figure 1 also situates the underpinning programme's theory of change, namely: theory of intervention (i.e. programme activities that produce outcome change); theory of change (i.e. developmental pathways in outcome changes); and implementation theory (i.e. implementation factors, like population characteristics and delivery style, that have an influence programme effectiveness). Full details of the programme content and implementation methods are available in the previously published trial protocol (O'Hare et al., 2018), which includes a TIDieR framework of programme description (Hoffmann et al., 2014) and a comprehensive programme logic model and programme theories. Further explanation of logic models and programme theory are available here (Connolly et al., 2017 chapter 2).

Figure 1 highlights how RR improves a teacher's ability to guide pupils through RR activities, which in turn improves pupils' reading comprehension metacognition (i.e. pupils explicitly applying appropriate reading strategies to support their comprehension) and reading accuracy. The programme's theory of change is that these improvements in reading comprehension metacognition and reading accuracy (secondary outcomes) result in improvements in pupil comprehension abilities that lead to further improvements in overall pupil reading ability and reading attainment (primary outcomes). It is a convention that primary outcomes are the key (primarily important) outcomes that the programme is intended to change. The secondary outcomes are theorized proximal, short-term changes that lead to distal change in the primary outcome of interest.

RESEARCH QUESTIONS

This study set out to investigate the effectiveness of FFT Literacy RR intervention and what influence the implementation choices had on the level of effectiveness by answering the following research questions:

1. Were there different effects of the universal or targeted RR intervention on the primary outcomes of comprehension and overall reading attainment in UK classrooms?
2. Were there different effects of the universal or targeted RR intervention on the secondary outcomes of reading accuracy and reading comprehension metacognition alongside the primary outcomes?
3. Did implementing the intervention with disadvantaged pupils have an impact on the level of effectiveness?

METHOD

The study consisted of two cluster RCTs that ran between September 2017 and June 2018. Ethical approval for the study was granted by the Research Ethics Committee of XYZ in the United Kingdom. The trial protocol has been published (O'Hare et al., 2018) and has been registered on the ISRCTN website and summarized here: (O'Hare et al., 2017).

Recruitment

Primary schools that met the following inclusion criteria were invited to participate in the trial:

1. Schools that had not previously participated in the RR programme.
2. Schools that were not included in any other ongoing EEF trial.
3. Schools that were clustered together within a geographical area with between 8–10 schools from each area, to maximize the probability of having at least three schools from each area randomized to deliver the intervention. Ensuring groups of schools were located within close proximity of one another facilitated CPD sessions such that schools did not have to travel far to receive training.

The project funder, the EEF, was and continues to be, focused on reducing the attainment gap between disadvantaged schools and more affluent schools. As a result, areas of high disadvantage were targeted to participate in this trial. The aim was that the participating schools had an overall mean of at least the national average of the school population ever in receipt of FSM an indicator of socio-economic disadvantage (29%). In practice, schools with higher-than-average levels of pupils eligible for FSM from the Northeast of England were contacted first and priority was given to these schools when they expressed an interest to participate during recruitment. In addition, other schools could express an interest to participate through the EEF website and were advised that schools with a higher-than-average level of FSM were being sought for participation.

Power

Sample-size calculations were carried out prior to school recruitment. Effect sizes for literacy interventions evaluated through a good quality RCT design would tend typically be in the range of .2–.3 (Borman et al., 2007; O'Hare et al., 2017; Tymms et al., 2011).

At the initial protocol stage, a power calculation for the universal intervention provided by Optimal Design software showed that 94 schools (clusters) would be needed to detect a significant effect size of Hedges g (g) = .2 if present with a power of .8.

The calculation used estimates as follows: Effect size = .2; p = .05; Intra-cluster correlation coefficient (ICC) = .14; r^2 = .50 (due to having a pre-test of New Group Reading Test – NGRT); and average cluster size n = 20.

Participants

We invited 142 schools to participate in the study. Of 119 schools agreeing to participate, five did not meet the inclusion criteria and 14 dropped out prior to pre-test. We chose to withdraw two further schools, after the pre-test and prior to randomisation, due to undisclosed non-compliance with the inclusion criteria (i.e. they were involved in other EEF trials). This left a total of 98 schools for pre-test and randomisation.

We asked participating schools to choose teachers and teaching assistants to be trained in basic FFT Literacy RR CPD. Following their training, they were invited to deliver the universal intervention to pupils in Year 4 (age 8 to 9) and the targeted intervention to pupils in Years 5 and 6 (age 9–11). We did this so we could compare the effectiveness of the two different interventions avoiding contamination across the same year group, that is, pupils having received both interventions. It was not feasible to recruit the additional schools required to conduct a multi-arm trial at the appropriate power required.

Our eligibility criteria for the targeted version required children (aged 9–11) to be selected for the programme on the basis that they were poor at reading comprehension with normal levels of decoding ability. Teachers selected these children before pre-baseline testing and pre-randomisation. This ensured that we were comparing equivalent pupils in the intervention and control groups (i.e. selected by teachers in all schools using the same approach).

We asked the teachers to select children in June 2017 using guidance and materials which were co-designed by FFT literacy and the evaluation team. These were based on a simple view of reading (Gough & Tunmer, 1986; Hoover & Gough, 1990). It was necessary for us to co-design a new process because there was no tool for quickly identifying the correct group of pupils available for use by teachers. The main outcome measure used in this trial (NGRT), as described in 'Outcomes and Measures' below, did not provide a measure of decoding for the majority of children and was therefore unsuitable for this selection process. It was also necessary to select children prior to pre-testing with the outcome measure as they were a sub-sample of whole classes.

Our co-designed selection guidance allowed teachers to compare each child in their class with two sets of criteria. The first set of criteria described the reading skills possessed by a child who could be classified as having normal-level decoding skills. The second set of criteria described difficulties that may be experienced by a child who struggles with reading comprehension. By comparing each child with these two sets of criteria, teachers were able to identify children who had a normal level of decoding skill but struggled with reading comprehension. Teachers in each school selected six pupils from each Years 5 and 6 class for the targeted intervention. Although there may have been more than six pupils who met the criteria for receiving the targeted intervention, teachers were asked to select only six.

The sessions for targeted pupils were delivered as an additional sessions rather than replacing other taught literacy sessions. However, to facilitate these sessions teachers were required to find the time during the school day which might have involved the children's withdrawal from other non-literary lessons or activities.

Randomisation

We performed randomisation after the collection of all baseline data and it was conducted at the school level using 'minimisation' to ensure the control and intervention groups were as evenly matched as possible (Torgerson & Torgerson, 2007). We performed minimisation stratification for several school-level covariates: baseline reading comprehension (NGRT passage comprehension score), disadvantage (% of school pupils ever in receipt of FSM) and baseline reading accuracy (NGRT sentence completion score). Minimisation was carried out using the QMinim package. Median values were calculated for reading comprehension score (NGRT passage comprehension), disadvantage and NGRT sentence completion

score. These medians were used to code schools as 'High' or 'Low' for each of these characteristics. These variables were then entered into QMinim for each school; and all variables were given a weight of one with the exception of reading comprehension which was double-weighted as an important predictor of the outcome of interest. The randomisation program then allocated schools to control or intervention groups in a way that minimized differences between groups for these three variables. Control schools acted as control group for both the universal and targeted interventions. Consequently, intervention schools included an intervention group for both the universal and targeted interventions. By assigning different year groups in the same schools to participate in different versions of the programme allowed evaluation of the two different versions of the programme (universal and targeted) within one sample of randomized schools.

Dosage

Dosage was defined as the total number of minutes of the programme calculated from minutes per week multiplied by the total number of weeks of delivery. Thirty two out of 49 intervention schools returned reported dosage data. Of those children in the universal RR programme, the mean delivery of minutes per week was 66.21 ($SD = 30.13$) with an average duration of the programme of 26.41 weeks ($SD = 5.06$) giving an average total number of minutes of 1748 min. Of those children in the targeted intervention group, the mean delivery of minutes per week of RR was 65.77 ($SD = 28.16$) and the average duration of the programme was 26.03 weeks ($SD = 6.31$) giving an average number of minutes of 1711 min. Both mean dosages are well more than the minimum recommended dosage.

Measures

The primary outcomes of the trial were two attainment indicators from the digital NGRT (GL Assessment, 2018): (1) Overall reading score; and (2) Comprehension subscale score: (see Table 1). The NGRT is an adaptive test which has high reliability (see Table 1 for the reliability of outcome measures) and measures both reading accuracy, comprehension and overall reading. GL Assessment reports high concurrent validity of the NGRT with correlations between the NGRT and teacher assessments of reading as $r = .81$ and Key Stage 2 SATs of $r = .75$ (GL Assessment, 2018).

All pupils who took the NGRT test received two subscales out of three possible subscales: sentence completion, passage comprehension and phonics. Sentence completion is the first subscale and was used as the secondary outcome reading accuracy. Pupils then progressed to the passage comprehension subscale (used as the primary outcome measure for reading comprehension attainment), unless they scored at an extremely low level for the sentence completion, in which case they progressed to the more basic phonics subscale instead. As described above, the number of items for each subscale is adaptive, as the difficulty of the next item is chosen based on their performance on items already administered. The maximum number of items depends on pupil performance. The aggregate overall reading score was calculated by GL before the researchers downloaded the scores from the GL software. The calculation was based on the two subscales each pupil received. All Y4, Y5 and Y6 pupils were provided with and asked to complete the NGRT digital test. The adaptive system meant that different test forms for different ages were not necessary. One caveat for the adaptive testing system used by the digital NGRT test was that reading comprehension (passage comprehension) outcome scores were not available for pupils who had scored at an extremely low level on the reading accuracy (sentence completion) subscale; these pupils received the phonics test instead. However, overall reading outcome data was still available for these pupils, as the GL system still generates an overall score, based on sentence completion and phonics elements. There were, therefore, fewer pupils with reading comprehension outcome data than there were pupils with overall reading data.

TABLE 1 Primary and secondary outcome measures.

	Measure	Level of measurement	Number of items	Reliability (Cronbach's alpha)
Primary outcomes				
Reading comprehension	New Group Reading Test – Passage Comprehension	Pupil	27	.9
Overall reading ability	New Group Reading Test – Total Score	Pupil	47	.9
Secondary outcomes				
Reading accuracy	New Group Reading Test – Sentence Completion	Pupil	20	.9
Reading comprehension metacognition	Assessment of Reading Comprehension Metacognition Scale (ARCMS)	Pupil	6	.7

The secondary outcomes of the trial (see Table 1) were reading accuracy scores from the sentence completion subtest of the NGRT and a pupil reading comprehension metacognition using the Assessment of Reading Comprehension Metacognition Scale (ARCMS; O'Hare et al., 2019) measure developed by the evaluation team. This researcher-created proximal measure was designed to closely reflect the principles of FFT Literacy RR and was developed to provide specificity matching to the intervention as no previously used standardized measures would be appropriate for this. Therefore, the evaluation team developed six questions that best matched the four key aspects of the programme: predicting, clarifying, questioning and summarizing. The six items related directly to the two overarching concepts of metacognition and comprehension. In this researcher-created proximal measure, the children were asked to 'think about what kinds of things you can do to help you to understand a story better when you read it' and included items related specifically to the strategies they were taught in the FFT Literacy RR sessions (e.g. 'Before I read a text I ask myself what I already know about the subject of the story (i.e., predicting))' and 'I find ways to help me understand words I am not sure about such as looking them up in a dictionary (i.e., clarifying).' The responses ranged from 1 (Always) to 5 (Never). Internal consistency reliability was measured using Cronbach's alpha and was deemed acceptable (.68) given that the conventional cut-off is .70 (Nunnally & Bernstein, 1994). The questions were designed by members of the evaluation team who had expertise in reading programmes to ensure the measure had content validity. Construct validity was assessed using Maximum Likelihood factor analysis as the data were normally distributed. One factor was extracted (eigenvalue 2.30) accounting for 26.46% of the variance. All items loaded .30 or above.

Data collection

We administered all tests within each school in classroom setting through a trained fieldworker with a class teacher in attendance. The children were each equipped with an individual computer or tablet, which was connected to a pair of headphones. We carried out the testing under exam conditions and the fieldworker delivered clear instructions to the children before the tests commenced. Children were not given additional support or guidance once the tests commenced. In addition, to the post testing, children completed the ARCMS through an on-line survey.

All children (in both the universal and targeted groups) completed a pre-test (NGRT form A) and a post-test (NGRT form B) on the outcome measures over the period of the evaluation (May/June 2017 and June/July 2018). The NGRT provides two versions of the digital tests, A and B, and a different test was used for pre- and post-testing to avoid potential practice effects. Of the children ($n = 4494$) who

completed both the pre- and post-test, $n = 3198$ were from the universal group and $n = 1296$ were from the targeted group. In addition, schools provided us with data on pupil names, date of Birth, and their Unique Pupil Number (UPN) which was matched with the National Pupil Database (NPD) to obtain accurate information on whether the pupil had ever received FSM which is a proxy for pupil disadvantage (Gorard & Siddiqui, 2019).

Statistical methods

Our analysis was conducted on an intention-to-treat basis and was carried out using STATA version IC15.1. We estimated the main effects of the intervention using multilevel mixed effects regression modelling to take account of the clustered nature of the data (clustered at the school level) and a series of models were estimated for each outcome (where pupil is level 1 and school is level 2). Mixed effects models account for both fixed (clustered at the school level) and random effects. Our analysis was not conducted at the pupils' individual class level. The analysis methodology, including detail on specific regression models and calculation of effect sizes, has been published in the peer-reviewed, pre-publication, analysis plan paper (O'Hare et al., 2018). As such, we have not deviated from our pre-published primary and secondary analysis reporting template. The reported results include raw scores for transparency of interpretation alongside effect sizes (Hedges g) and statistical significance. Data were archived and made available to an external and independent analysis team for the purposes of verifying the primary analysis.

Firstly, two models were conducted for the universal intervention with the NGRT comprehension score and overall reading score forming the dependent variable of each model and the independent variables including a dummy variable representing whether the child was a member of the intervention or control group (coded '1' and '0' respectively) and pupils' baseline scores at pre-test. We repeated the same analysis for the targeted intervention. Each primary analysis model was conducted using standardized pre-test scores as a covariate (z-scores) to control for variation in pre-test scores between groups (O'Hare et al., 2017).

Those schools that dropped out of delivering the intervention were encouraged to allow post-testing. Following the intention-to-treat methodology, schools and participants that were lost-to-follow up were included in as many analyses as possible. The pre-test data from drop-out schools was still included in raw means and the maximum number of participants was used for each analysis model, that is, pupils were included in our analysis Model 1 even if they were missing from analysis Model 2.

Our trial protocol specified that if missing data was higher than 5%, a 'missing at random' analysis would be carried out to determine if multiple imputation was required. This was carried out by running chi-squared analyses of group allocation and missingness (missingness was a dummy variable, coding '1' for missing outcome data and '0' for not missing). Multiple imputation was performed using chained iterations to fill in missing values in multiple variables using univariate imputation with fully conditional specification of prediction equations. Twenty imputations were carried out to reduce the risk of the 'Monte Carlo' error (simulation error). Twenty imputations are recommended for between 10% and 30% of missing data (Graham et al., 2007). A total of 200 iterations (with a burn-in of 10) were carried out, and estimates were combined using Rubin's pooling rules. This then allowed the primary analysis to be repeated using the imputed data, and the difference in primary effects was compared with the complete case analysis to investigate if the presence of missing data impacted the results. Levels of missing data for each outcome and complete case analysis are reported in the 'Sensitivity analysis: multiple imputation' section of the findings.

RESULTS

Participant flow through the study

Participant flow is detailed in Figures 2 and 3. One hundred schools were pretested and randomized (51 intervention and 49 control). Two of these intervention schools were excluded after

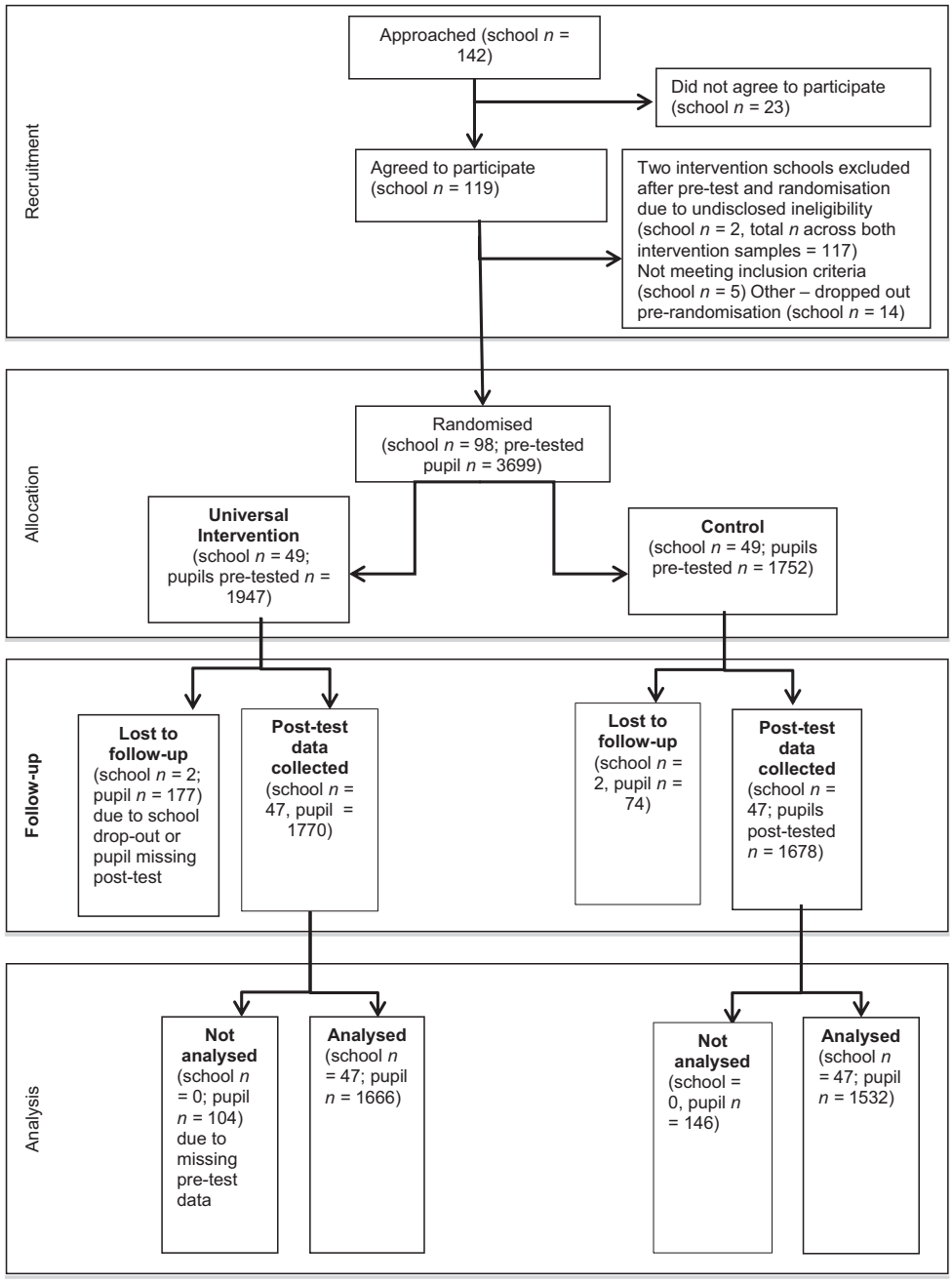

FIGURE 2 Participant flow: Universal intervention.

pretesting and randomisation as they were ineligible for the trial but had not disclosed this on their signed memorandum of understanding. This meant that 98 eligible schools had been pretested and randomized. In the universal intervention sample, $n = 3699$ pupils ($n = 98$ schools) were pretested. The exclusion of two schools = occurred post-randomisation. At the point of analysis, there were $n = 3198$ pupils. This is an overall attrition rate of 13.5% for the universal intervention sample. Attrition for intervention group was 14.43% and attrition for the control group was 12.56%. In the

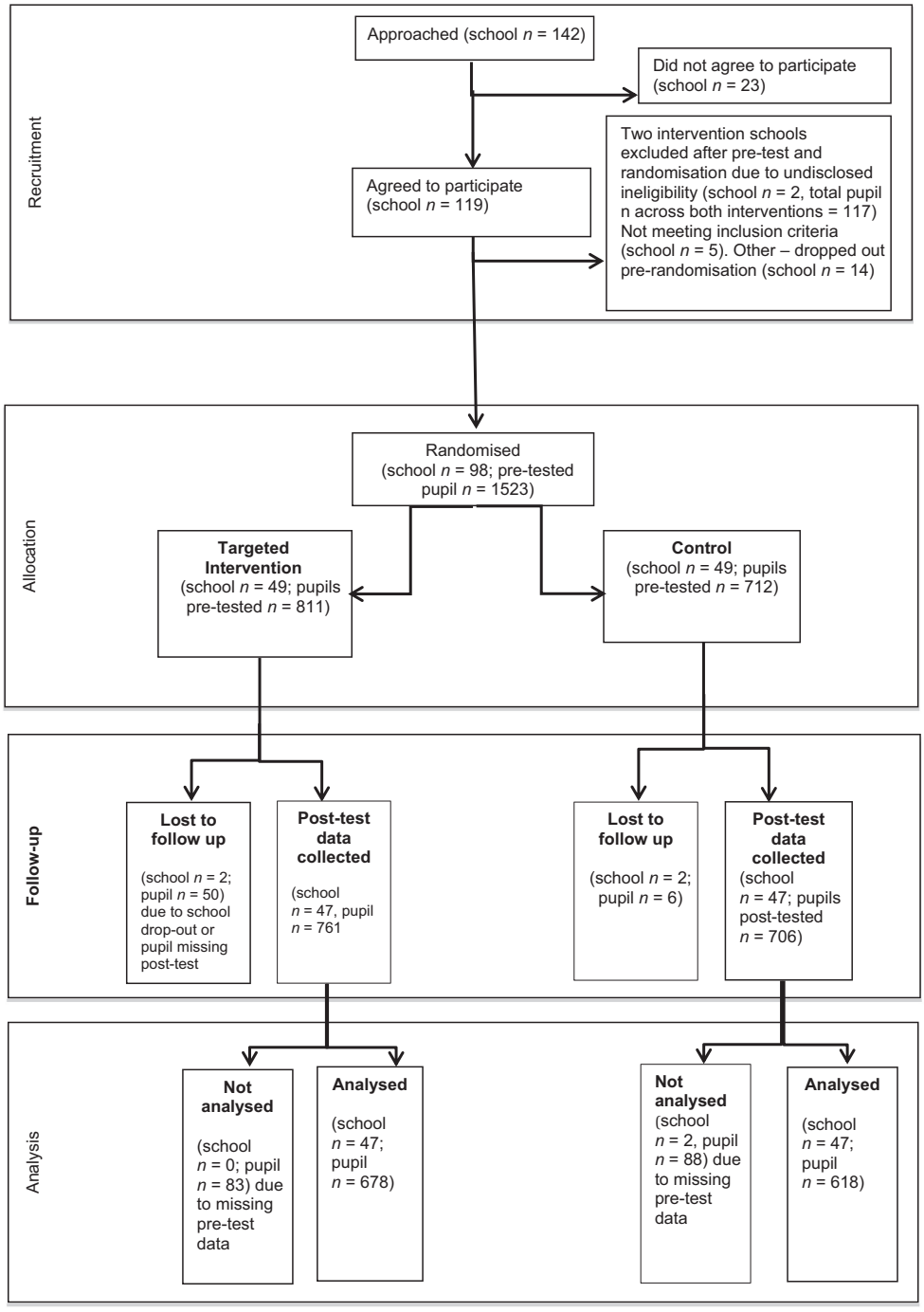

FIGURE 3 Participant flow: Targeted intervention.

targeted intervention sample, $n = 1523$ pupils ($n = 98$ schools) were pretested. At the point of analysis, there were $n = 1296$ pupils. This is an overall attrition rate of 14.9% for the targeted intervention sample. In the targeted intervention sample, attrition for the intervention group was 16.4% and attrition for the control group was 13.2%. Any pupils missing from post-test who were from schools

that remained active in the trial were not present for post-test, either due to absence or leaving the school. The analysed *N* is lower than the baseline *N*, as pupils needed both pre-test and post-test scores to be included in the complete case analysis models (both these variables are included in the primary models). Data withdrawal forms were issued before pretesting, and schools were instructed not to pre-test pupils who had withdrawn from the data collection. The pre-test data collection was the point at which the evaluation team first received any pupil-level data from schools. The number of pupil withdrawals is, therefore, not calculated. The analysed N refers to the number of pupils who the evaluation team received both pre-test and post-tests for (i.e. who sat both NGRT pre-tests and post-tests resulting in an overall reading score for both).

Baseline data

Table 2 displays the baseline data for the universal Year 4 programme evaluation. It shows the school and pupil baseline data for the intervention and control groups. School level characteristics are broken down by Ofsted rating (an independent rating of school quality) and school type. Pupil characteristics include number of pupils per school, % eligibility for FSM (indicator of disadvantage) and pre-test scores on primary outcome measures. Analysis of pre-test differences on outcomes show that the effect sizes are very small between control and intervention schools. This means that the groups did not differ strongly for reading scores before the intervention was delivered.

Table 3 shows the baseline data for the targeted Years 5 and 6 programme evaluation. It is noted that there are small non-significant effect sizes for the difference between control and intervention schools in the targeted sample at baseline on the NGRT pre-test for primary outcomes. Effect sizes of .2 or over would indicate that the sample groups were substantively different, but it should be noted that these effect sizes were found to be in the within a range of .15 to .18 for the three reading score variables. This means that prior to the intervention being delivered, the intervention schools were already performing at a slightly higher level on the reading outcomes than the control schools. However, the effect sizes calculated for the primary analysis later in the report are calculated in a way that controls for imbalance at pre-test. Where pre-test scores were included as a covariate in the analysis models, they were standardized to have a mean of zero, to control for pre-test imbalances. The imbalance at baseline, therefore, may have meant that the intervention was being delivered in a group of schools who had slightly higher reading attainment, but it does not strongly impact the interpretation of the results at post-test.

Outcomes and estimation

Overall, the targeted version of FFT Literacy RR programme had significant positive effects on both primary and secondary reading outcomes, as shown in Table 4. Specifically, the targeted version of the programme improved primary outcomes of overall reading (effect size $g = .14$) and reading comprehension (effect size $g = .18$) However, the universal version of the programme showed no significant improvements in terms of the primary outcomes with no significant effects on either overall pupil reading or pupil reading comprehension (Table 4). The school-level intra-cluster correlation coefficient (ICC) was calculated at the analysis stage for the universal sample as ICC = .04, and as ICC = .18 for the targeted sample.

Table 5 shows that the targeted programme improved the secondary outcomes of reading accuracy and pupil reading comprehension meta-cognition. There was one significant positive effect of the universal version on secondary reading outcomes, that is, improved pupil reading comprehension meta-cognition but there was no significant effect on pupil reading accuracy (Table 5).

TABLE 2 Balance at baseline in intervention and control groups: universal intervention.

	Intervention group		Control group		
School-level (categorical)	***n/N* (missing)**	**Count (%)**	***n/N* (missing)**	**Count (%)**	
Ofsted rating:	49 (0)		49 (0)		
Outstanding		14.3		8.2	
Good		69.4		83.7	
Satisfactory		0		0	
Inadequate		0		2	
Requires improvement		16.3		6.1	
Academy (converter)	4	8.2	8	16.33	
Academy (sponsor led)	4	8.2	2	4.08%	
Community school	25	51	23	46.94	
Faith school (academy converter)	1	2	1	2.04	
Faith school (foundation school)	0	0.00	1	2.04	
Faith school (voluntary aided)	7	14.3	7	14.29	
Faith school (voluntary controlled)	3	6.1	3	6.12	
Foundation school	4	8.2	4	8.16	
Free school	1	2	0	.00	
School-level (continuous)	Mean		Mean		
Y4 Pupil *N* per school at baseline	39		36		
Pupil-level (categorical)	*N*	Count (%)	*N*	Count (%)	
Eligible for FSM	348	18.34	418	20.38	
Pupil-level (continuous)	*n* (missing)	Mean (*SD*)	*n* (missing)	Mean (*SD*)	Effect size
Pre-test NGRT overall reading scale (primary outcome)	1947 (104)	248.54 (62.72)	1752 (146)	246.10 (64.64)	.04 (CI: −.03 to .10)
Pre-test NGRT reading comprehension (primary outcome)	1832 (219)	248.61 (61.39)	1619 (279)	247 (62.24)	0.03 (CI: −.04 to .09)

Sensitivity analysis: Multiple imputation

The proportion of missing data at pre-test and post-test was greater than 5% for both primary outcomes (see Table 6). The control group showed higher numbers of missing data for pre-test overall reading ($p < .01$) and for pre-test reading comprehension ($p < .01$). Intervention showed higher numbers of missing data for post-test overall reading ($p < .01$). This pattern in missingness means that data was presumed to be missing at random.

As per trial protocol, multiple imputation analysis was then carried out comparing primary analysis with imputed data and complete cases data. The pattern of results for primary analysis using the two versions of the data set was the same, that is, all significant and non-significant results remained the same. This analysis of the data suggests that missing data do not negatively impact the strength of the evidence for efficacy of the targeted intervention or the evidence of a lack of efficacy for the universal intervention.

TABLE 3 Balance at baseline in intervention and control groups: targeted intervention.

	Intervention group		Control group		
School-level (categorical)	***N* (missing)**	**Count (%)**	***N* (missing)**	**Count (%)**	
Ofsted rating:	49 (0)		49 (0)		
Outstanding		14.3		8.2	
Good		69.4		83.7	
Satisfactory		0		0	
Inadequate		0		2	
Requires improvement		16.3		6.1	
Academy (converter)	4	8.2	8	16.33	
Academy (sponsor led)	4	8.2	2	4.08	
Community school	25	51	23	46.94	
Faith school (academy converter)	1	2	1	2.04	
Faith school (foundation school)	0	0	1	2.04	
Faith school (voluntary aided)	7	14.3	7	14.29	
Faith school (voluntary controlled)	3	6.1	3	6.12	
Foundation school	4	8.2	4	8.16	
Free school	1	2	0	.00	
School-level (continuous)	Mean		Mean		
Y5 & Y6 pupil *N* per school at baseline	16		15		
Pupil-level (categorical)	*N*	Count (%)	*N*	Count (%)	
Eligible for FSM	158	17.67	182	22.75	
Pupil-level (continuous)	*n* (missing)	Mean (*SD*)	*n* (missing)	Mean (*SD*)	Effect Size
Pre-test NGRT overall reading Scale (primary outcome)	811 (83)	277.05 (45.30)	712 (88)	268.57 (47.30)	.18 (CI: .08–.28)
Pre-test NGRT reading comprehension (primary outcome)	804 (90)	270.57 (53.59)	704 (96)	262.70 (52.33)	.15 (CI: .05–.25)
Pre-test NGRT reading accuracy (secondary outcome)	811 (83)	289.76 (43.16)	712 (88)	281.59 (46.415)	.18 (CI: .08–.28)

Subgroup analysis for disadvantaged pupils

The primary analyses were repeated for disadvantaged pupils, that is, only those in that have ever received FSM.

Table 7 shows a significant effect of the universal intervention was found for Overall Reading Ability for FSM pupils (effect size $g = .08$). This means that, although there was no effect of the Universal intervention for the overall sample there was a small positive effect of the universal intervention for FSM pupils. However, no significant effect of the universal intervention was found for Reading Comprehension for FSM pupils.

TABLE 4 Evaluation of effects on primary outcomes of reading ability and reading comprehension for both universal and targeted programmes: Complete case analysis.

Outcome		Raw means					Effect size	
		Treatment group		Control group		*N* in model (intervention; control)		
Primary outcome	**Covariate**	***n* (missing)**	**Mean Post-test [*SD*]**	***n* (missing)**	**Mean Post-test [*SD*]**		**Hedges *g* [95% CI]**	***p***
Universal version on primary outcomes								
NGRT overall reading ability	Pre-test NGRT overall reading ability	1770 (281)	284.98 [57.28]	1678 (220)	281.72 [61.81]	3198 (1666, 1532)	.00 [−.06, .07]	.795
NGRT reading comprehension	Pre-test NGRT reading comprehension	1722 (329)	288.06 [57.64]	1605 (293)	288.52 [57.13]	2958 (1557, 1401)	−.02 [−.09, .06]	.494
Targeted version on primary outcomes								
NGRT overall reading score	Pre-test NGRT overall reading score	761 (133)	309.71 [47.16]	706 (94)	288.46 [49.98]	1296 (678, 618)	.14 [.04, .25]	<.001
NGRT Reading Comprehension	Pre-test NGRT Reading Comprehension	755 (139)	308.51 [51.38]	693 (107)	285.71 [52.06]	1270 (668, 602)	.18 [.07, .29]	<.001

TABLE 5 Evaluation of effects on secondary outcomes of reading accuracy and pupil reading comprehension meta-cognition for both universal and targeted programmes.

NGRT reading accuracy	Coef.	Std. err.	*z*	Significance	95% CI	
Universal version on reading accuracy						
Group	2.21	2.26	.98	.33	−2.22	6.65
Pre-test NGRT reading accuracy	36.43	.76	47.68	<.001	34.93	37.92
Constant	289.32	1.63	177.90	<.001	286.13	292.50
Universal version on pupil comprehension meta-cognition						
Group	1.19	.31	3.84	<.001	.58	1.79
Pre-test ARCMS reading comprehension meta-cognition	.06	.09	.61	.55	−.13	.24
Constant	18.18	.22	82.39	<.001	17.74	18.61
Targeted version on reading accuracy						
Group	8.57	4.21	2.04	.04	.32	16.82
Pre-test NGRT reading accuracy	23.62	1.61	14.70	<.001	20.47	26.77
Constant	293.48	3.06	96.02	<.001	287.49	299.47
Targeted version and pupil comprehension meta-cognition (ARCMS)						
Group	1.13	.37	3.08	<.001	.41	1.85
Pre-test ARCMS reading comprehension	−.25	.17	−1.50	.13	−.58	.08
Constant	18.43	.26	71.12	<.001	17.92	18.94

TABLE 6 Missing data for each primary outcome scale for universal and targeted samples.

Sample	Outcome scale	% missing
Universal	Pre-test overall reading	6.33
	Post-test overall reading	12.68
	Pre-test comprehension	12.61
	Post-test comprehension	16.03
Targeted	Pre-test overall reading	10.09
	Post-test overall reading	13.40
	Pre-test comprehension	10.98
	Post-test comprehension	14.52

Table 8 shows significant positive effects of the Targeted Intervention was found on both overall reading (ES = .2) and reading comprehension (ES = .25) for FSM pupils. This shows that disadvantaged pupils in this study benefitted from the targeted intervention in terms of both overall reading and reading comprehension. These effect sizes are bigger in both cases than in the full population seen in Table 4.

Replication and verification of primary analysis

Primary outcome data was shared with an external and independent analysis team from another institution, who re-analysed the data and verified the results reported here. The replication analysis was

TABLE 7 Evaluation of effects on primary outcomes of reading ability and reading comprehension of universal programme for FSM subgroup.

Outcome		Raw means					Effect size	
		Treatment group		Control group		*N* in model (intervention; control)		
Primary outcome	**Covariate**	***n*** **(missing)**	**Mean Post-test [*SD*]**	***n*** **(missing)**	**Mean Post-test [*SD*]**		**Hedges *g* [95% CI]**	***p***
NGRT overall reading ability	Pre-test NGRT overall reading ability	618 (127)	268.42 [56.39]	593 (112)	259.96 [66.23]	1120 (572; 548)	.08 [−.04, .2]	.05
NGRT reading comprehension	Pre-test NGRT reading comprehension	594 (151)	270.76 [57.06]	547 (158)	270.76 [52.09]	989 (512; 477)	−.05 [−.17, .08]	.37

TABLE 8 Evaluation of effects on primary outcomes of reading ability and reading comprehension of targeted programme for FSM subgroup.

Outcome		Raw means					Effect size	
		Treatment group		Control group		*N* in model (intervention; control)		
Primary outcome	**Covariate**	***n* (missing)**	**Mean Post-test [*SD*]**	***n* (missing)**	**Mean Post-test [*SD*]**		**Hedges *g* [95% CI]**	***p***
NGRT Overall reading score	Pre-test NGRT Overall reading score	290 (52)	302.72 [47.10]	305 (54)	282.54 [49.70]	543 (280; 263)	.2 [.03, .37]	.006
NGRT Reading Comprehension	Pre-test NGRT Reading Comprehension	286 (56)	302.37 [48.43]	299 (60)	279.68 [50.93]	529 (273; 256)	.25 [.08, .42]	<.001

commissioned by the funder (Education Endowment Foundation – EEF) for internal checks and decision-making before regranting the intervention. The analysis is not publicly available but may be available on request to the EEF.

DISCUSSION

Looking back at the logic model (Figure 1) the results provide evidence supporting the programme's theory of intervention (i.e. programme activities are producing outcome change) for the targeted version of FFT Literacy RR. That is, RR improves reading comprehension and reading ability of Years 5 and 6 pupils (aged 9–11) in UK schools. However, the findings do not suggest that the universal version of the programme for Year 4 (age 8 to 9) pupils provided the same benefits.

So why did the targeted version have better effects than the universal version? From the data and findings, it is not possible to attribute an experimental or causal explanation for why the targeted version worked and the universal version did not, as they differed in more ways than targeted pupil selection on comprehension deficits for example, age of pupils and size of delivery groups were also different across the two interventions. In reality, this is two different trials (with similar methods) rather than two arms being compared within a single trial. However, there is some correlational evidence from the study along with theory and literature that can help us generate some hypotheses (not causality) for the different results. The pupils in the targeted version were specifically chosen as those who best aligned with the programme theory of intervention. That is, it helps develop comprehension ability in those who are typically developing in reading accuracy/decoding but below average in reading comprehension. This notion is also supported by the fact that the significant effect on comprehension was slightly larger (effect size $g = .18$) than on overall reading (effect size $g = .14$) within the targeted group.

Another hypothesis could be that the older age of the readers 9–11 (Years 5 and 6) in the targeted intervention were better able to understand and apply the four core concepts of the programme (predicting, clarifying, questioning and summarizing) than the younger readers in the universal version aged 8–9 (Year 4). This would concur with the previous research on younger students being less able to engage with meta-cognitive or strategic thinking about reading (Duffy, 1993; Hacker & Tenent, 2002).

A common explanation for variation in reading intervention success is variation in dosage, that is, the higher the dosage, the more effective the intervention. In this study, however, the universal intervention was delivered at a higher dosage but was less effective than the targeted intervention. Although the recommended dosage of the targeted intervention is higher than the recommended dosage of the universal intervention, this does not mean that any deviation from this makes a school non-compliant with the intervention. The much higher-than-minimum dosage found in all intervention schools underlines how feasible and successfully implemented the programme was in this study. Again, we cannot be sure that the minimum effective dosage declared by the developers is in fact 'effective', and the minimum effective dosage remains untested due to the high dosage in this study.

The findings also provide some evidence to support the theory of change within the logic model (i.e. developmental pathways between outcome changes) in that there were improvements in both the proximal secondary outcomes of reading comprehension metacognition and improved reading accuracy and distal primary outcomes of reading comprehension and overall reading attainment. This chain of outcome change is still a working hypothesis and, again, further analysis could be helpful to strengthen this theory of change.

The sub-group analysis suggested that the benefits of RR are even greater for disadvantaged pupils. Even the universal version provided some significant benefits to disadvantaged pupils. These findings suggest that socio-economic disadvantage could be considered in the selection process for targeted intervention. So, in addition to selecting pupils based on teacher's perceptions of pupils' poor reading comprehension ability and normal decoding skills, the teachers could also use disadvantage as an additional variable for selecting pupils to receive the targeted programme. Inclusion of disadvantage

could improve the reliability and validity of the targeting process as pupils selected solely on teacher perceptions of pupil reading accuracy and comprehension is open to teacher perception bias or error. One caveat to this is, it may be that the disadvantage benefits of the programme are acting at the school level and therefore, selecting schools for the intervention based on higher levels of disadvantage (rather than individual pupils) is enough to confer these additional benefits. Generally, more study is required to clarify the benefits of RR for addressing disadvantages and inequality in outcomes.

So overall, this evaluation provides strong support for further implementation of the targeted version of FFT Literacy RR intervention. The main reasons being the consistent evidence of its effect on reading outcomes; its focus on pupils who benefit most through a targeting process; its potential for reducing the attainment gap between disadvantaged and non-disadvantaged pupils; the lack of strong evidence for other reading comprehension interventions; and it is low cost because it uses a teacher professional development model.

Limitations

The generalisability of these findings is quite good for pupils of similar characteristics to the sample tested, that is, Years 4, 5 and 6 pupils in English primary schools with above average levels of pupils in receipt of FSM, due to the scale and size of this study. However, generalisability to other age groups may be less certain.

A limitation of this study is that there was some potential for teacher selection bias in their identification of pupils in receipt of the targeted version of the programme. However, both control and intervention teachers did make their selections before randomisation and their choices were based on standardized guidance based on the simple view of reading (Gough & Tunmer, 1986; Hoover & Gough, 1990). It could be argued that an objective method with evidence of reliability and validity is required. There is merit in this argument and there is a need for more research on the selection method. However, there is an overriding practical concern that if the programme is to be scaled up and available to more teachers, they need a simple method and does not involve complicated selection instruments.

A threat to internal validity is the risk associated with children joining the trial after pre-testing and randomisation, thus introducing group imbalance beyond that at the point of pre-testing. The number of children who were missing from the pre-test and randomisation data was $n = 104$ and $n = 146$ for intervention and control in the universal intervention sample and $n = 83$ and $n = 88$ for the targeted intervention sample.

A limitation of the analysis is that class-level clustering was not included in the multi-level models. We collected data at the school level via digital NGRT testing but data was not gathered on class (for the Universal intervention) or for targeted delivery group (for the Targeted intervention). Considering the intervention was delivered in a clustered way, below school level, that is, class or targeted group it is possible that variance would have been observed between teachers/teaching assistants who delivered the intervention. Secondly, this means that class or group-level peer effects are not considered. This must be addressed in future evaluations of Reciprocal Reading and reduces the explanatory power of the present analysis.

Another potential issue with this evaluation was that the NGRT test uses adaptive testing. Therefore, pupils who perform very poorly on accuracy questions may not get access to the comprehension questions, although this may not have impacted upon the pupils in the targeted group as they were selected on normal accuracy ability. Also, distributions of the pre-test and post-test of primary outcomes showed no floor or ceiling effects.

One further issue about measures is that the secondary outcome measures were researcher-created proximal measures as no existing tests had the specificity to measure these outcomes (e.g. pupil comprehension meta-cognition). However, these tests were designed with scrutiny of content validity and post hoc tests of reliability and validity were conducted. In addition, these secondary measures are

mainly to explore the theory of change rather than effectiveness of the intervention. Effectiveness of the programme should mainly rest on the performance of the intervention on the primary standardized outcome measures of comprehension and overall reading attainment.

A limitation to the dosage analysis is that it is possible that high-dosage schools were more likely to submit dosage data, thus under-representing schools with low-dosage in the analysis and interpretation.

Finally, the sample sizes for the subgroup analyses are comparatively smaller than the primary analyses in the main report, but still of a substantial size. However, the interpretation of the results should take sample size into consideration.

CONCLUSION

In conclusion, it is an important finding that this study demonstrated an effective application of RR instruction. This importance is enhanced due to the proportion of language and literacy-focused interventions rigorously evaluated in the United Kingdom not showing significant effects on reading outcomes (Education Endowment Foundation, 2018). Furthermore, with the Department for Education (2013b) in the UK advocating a universal focus on comprehension instruction, a targeted version of FFT Literacy RR can offer additional support to those who may not benefit from the curriculum's universal approach.

Finally, this evaluation shows that even if an intervention is underpinned by evidence-based practice and has a strong theory of change, implementation factors such as appropriate targeting and the developmental stage of participants can still determine the effectiveness of reading comprehension interventions.

AUTHOR CONTRIBUTIONS

Liam O'Hare: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Patrick Stark:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Maria Cockerill:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Katrina Lloyd:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Shelia McConnellogue:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Aideen Gildea:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Andy Biggart:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing. **Christine Bower:** Data curation; formal analysis; investigation; methodology; project administration; writing – original draft; writing – review and editing. **Paul Connolly:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions and participation of all the schools, teachers and pupils involved in this study. They also gratefully acknowledge the funding provided by the Education Endowment Foundation to conduct the study.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Liam O'Hare https://orcid.org/0000-0002-4453-2880

TWITTER

Liam O'Hare @lohare

REFERENCES

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal*, *44*(3), 701–731. https://doi.org/10.3102/0002831207306743

Chiu, C. W. T. (1998). *Synthesizing metacognitive interventions: What training characteristics can improve reading performance? [annual meeting presentation]*. American Educational Research Association annual meeting, San Diego, CA.r.

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using Randomised Controlled Trials in Education*. SAGE.

De Corte, E., Verschaffel, L., & Van De Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. *British Journal of Educational Psychology*, *71*(4), 531–559. https://doi.org/10.1348/000709901158668

Department for Education. (2013a). *English programmes of study: key stages 1 and 2 National curriculum in England*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/335186/PRIMARY_national_curriculum_-_English_220714.pdf

Department for Education. (2013b). *The national curriculum in England: Framework document*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381344/Master_final_national_curriculum_28_Nov.pdf

Duffy, G. G. (1993). Teachers' progress toward becoming expert strategy teachers. *The Elementary School Journal*, *94*(2), 109–120. https://doi.org/10.1086/461754

Education Endowment Foundation. (2018). *Completed projects*. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/reports/

Education Endowment Foundation. (2022a). *Sutton trust-education endowment foundation teaching and learning toolkit*. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit

Education Endowment Foundation. (2022b). *Reading comprehension strategies*. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/reading-comprehension-strategies

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393–447. https://doi.org/10.3102/00346543071003393

GL Assessment. (2018). *Technical information*. New Group Reading Test (NGRT) Digital Edition. https://support.gl-assessment.co.uk/knowledge-base/assessments/ngrt-support/general-information/technical-guidance/

Gorard, S., & Siddiqui, N. (2019). How trajectories of disadvantage help explain school attainment. *SAGE Open*, *9*(1), 215824401882517. https://doi.org/10.1177/2158244018825171

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

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Hacker, D. J., & Tenent, A. (2002). Implementing reciprocal teaching in the classroom: Overcoming obstacles and making modifications. *Journal of Educational Psychology*, *94*(4), 699–718. https://doi.org/10.1037/0022-0663.94.4.699

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, *348*, g1687. https://doi.org/10.1136/bmj.g1687

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. https://doi.org/10.1007/BF00401799

Kispal, A. (2008). *Effective teaching of inference skills for children: Literature review. Research Report DCSF-RR031*. National Foundation for Educational Research. https://www.nfer.ac.uk/publications/edr01/edr01.pdf

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014). *Summer active reading programme: evaluation report and executive summary*. Education Endowment Foundation.

McPhillips, M., & Sheehy, N. (2004). Prevalence of persistent primary reflexes and motor problems in children with reading difficulties. *Dyslexia*, *10*(4), 316–338. https://doi.org/10.1002/dys.282

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development.

Nunnally, J., & Bernstein, L. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill Higher.

O'Hare, L., Lloyd, K., Stark, P., McConnellogue, S., Biggart, A., & Cockerill, M. (2017). *Reciprocal Reading: A training programme for teachers aimed at improving reading skills of pupils*. ISRCTN Trial Registry. https://www.isrctn.com/ISRCTN81582662

O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., Biggart, A., Connolly, P., & Bower, C. (2019). *Reciprocal reading: Evaluation report*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Reciprocal_Reading.pdf

O'Hare, L., Stark, P., McConnellogue, S., Lloyd, K., Cockerill, M., & Biggart, A. (2018). Protocol: A cluster randomised controlled trial of Reciprocal Reading: A teacher training comprehension programme. *International Journal of Educational Research*, *92*, 30–42.

Okkinga, M., van Steensel, A., van Gelderen, A., & van Schooten, E. (2018). Effectiveness of reading strategy intervention in whole classrooms: A meta-analysis. *Educational Psychology Review*, *30*(4), 1215–1239. https://doi.org/10.1007/s10648-018-9445-7

Palincsar, A. S. (2013). Reciprocal teaching. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 369–371). Routledge.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117–175. https://doi.org/10.1207/s1532690xci0102_1

Pressley, M. (2002). *Reading instruction that works: The case for balanced teaching* (4th ed.). Guilford Press.

Rockets, R. (2023). *How to use reciprocal reading*. https://www.readingrockets.org/strategies/reciprocal_teaching

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

Sencibaugh, J. M. (2007). Meta-analysis of reading comprehension interventions for students with learning disabilities: Strategies and implications. *Reading Improvement*, *44*(10), 6–22.

Swanson, H. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities*, *32*(6), 504–532. https://doi.org/10.1177/002221949903200605

Thurston, A., Roseth, C., O'Hare, L., Davison, J., & Stark, P. (2016). *Talk of the town. Evaluation report and executive summary*. Education Endowment Foundation.

Torgerson, C. J., & Torgerson, D. J. (2007). The use of minimization to form comparison groups in educational research. *Educational Studies*, *33*(3), 333–337. https://doi.org/10.1080/03055690701423267

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., & Miller, D. (2011). Improving attainment across a whole district: School reform through peer tutoring in a randomized controlled trial. *School Effectiveness and School Improvement*, *22*(3), 265–289. https://doi.org/10.1080/09243453.2011.589859

How to cite this article: O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., Biggart, A., Bower, C., & Connolly, P. (2023). Comparing the effectiveness of two reciprocal reading comprehension interventions for primary school pupils in disadvantaged schools. *British Journal of Educational Psychology*, *00*, e12623. https://doi.org/10.1111/bjep.12623