



Education
Endowment
Foundation

LIT Programme

Evaluation Report and Executive Summary

October 2014

Independent evaluators:



NatCen
Social Research that works for society

Claire Crawford, Institute for Fiscal Studies

Amy Skipp, NatCen Social Research

The Education Endowment Foundation (EEF)



The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- Identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- Evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale;
- Encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust, as lead charity, in partnership with Impetus Trust (now part of Impetus – The Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



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

James Richardson

Senior Analyst
Education Endowment Foundation
9th Floor, Millbank Tower
21-24 Millbank
London
SW1P 4QP

p: 020 7802 1924

e: james.richardson@eefoundation.org.uk

w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the Institute for Fiscal Studies (IFS) and NatCen Social Research. Claire Crawford, Research Fellow at IFS and Assistant Professor of Economics at the University of Warwick, led the impact evaluation. Amy Skipp, Research Director of the Children and Young People team at NatCen Social Research, was responsible for overseeing the pupil testing and leading the process evaluation.

Contact details:

Claire Crawford
Research Fellow, Institute for Fiscal Studies

7 Ridgmount Street
London
WC1E 7AE

p: 020 7291 4800
e: claire_c@ifs.org.uk

Amy Skipp
Research Director, NatCen Social Research

35 Northampton Square
London
EC1V 0AX

p: 020 7549 8502
e: amy.skipp@natcen.ac.uk

Contents

Executive Summary	2
Introduction.....	5
Methodology	8
Impact Evaluation	17
Process Evaluation	28
Conclusion	38
References	42
Appendices	44

Executive summary

The project

The Literacy Intervention Toolkit (LIT) programme aims to improve the reading ability of children in Year 7 who scored below Level 4 at the end of primary school using a method known as reciprocal teaching. Reciprocal teaching methods encourage children to 'become the teacher'. They are taught how to apply four comprehension strategies: summarising, clarifying, questioning, and predicting. These strategies enable children to check that they understand the content of the material they are reading and can make inferences based on what they have read.

The LIT programme is tightly structured, providing training to staff as well as a set of detailed lesson plans on the use of reciprocal teaching to deliver basic instruction in literacy. However, the method of delivery is not particularly prescriptive. It can be used to teach whole classes or small groups, can be delivered by teachers or teaching assistants, and can be offered in addition to or instead of regular English classes. In this evaluation, children typically received 3–4 hours of LIT tuition per week for eight months, mostly delivered in small groups.

The programme was devised by the Hackney Learning Trust, who delivered the training sessions for staff and developed a detailed set of lesson plans for them to follow.

The primary outcome was reading ability as assessed by scores on the Access Reading Test (ART).

The LIT programme was tested using a randomized control trial (RCT). 41 schools were recruited, with 22 schools randomly allocated to a treatment group and 19 schools to a control group.


Key conclusions:

1. This evaluation cannot conclude with certainty what impact the LIT programme had on reading ability for those pupils who received the intervention.
2. This evaluation did not find evidence of a significant impact on reading ability at the year group level (comparing *all* Year 7 pupils in treatment schools with similar Year 7 pupils in control schools.).
3. Teachers felt that the programme facilitated 'healthy debate' within the classroom, increased confidence in pupils who struggled with core literacy skills, and promoted independent learning. However, they did not feel it worked well with children with underlying cognitive issues requiring intensive vocabulary support.
4. Feedback from those who used the programme suggested groups of 5–6 children led by qualified teachers or teaching assistants with experience of delivering literacy interventions worked best.

What impact did it have?

The primary objective of the independent evaluation was to test whether the LIT programme had a significant impact on the literacy skills of the approximately 15% of Year 7 children who were eligible to take part (those with the poorest literacy skills). Unfortunately, it was not possible to do this because the characteristics of pupils in treatment and control schools were too different to yield an unbiased estimate of the impact of the programme. However, it was possible to compare *all* Year 7 pupils in treatment schools with similar Year 7 pupils in control schools, regardless of whether they received the LIT programme. This captures the effect of the LIT programme on the whole year group.

These results suggest that the reading scores of Year 7 pupils in schools running the LIT programme rose 0.09 standard deviations more over the course of the year than those of similar Year 7 pupils in control schools. This effect size of 0.09 is equivalent to 1 month's additional progress in reading. However, statistical tests indicate that this estimate is not 'significant', i.e., we cannot be sure that the difference in scores is not due to chance. It is important to note that this is *the average effect across all pupils in Year 7*, i.e., across the 15% of pupils who received the LIT programme as well as the 85% of pupils who did not. It is possible that the effect on those who received it was much higher.

Group	No. of pupils	Effect size (95% confidence interval)*	Estimated months' progress	Is this finding statistically significant?*	Evidence strength**	Cost of approach***
All pupils	4,413	0.09 (-0.04, 0.22)	+1	No		£
<p>* Effect sizes with confidence intervals that pass through 0 are not 'statistically significant', suggesting that the difference may have occurred by chance.</p> <p>** For more information about evidence ratings, see Appendix D</p> <p>*** For more information about cost ratings, see Appendix F in the main evaluation report.</p>						

How secure is this finding?

The intention was to estimate the impact of the LIT programme by comparing the subset of pupils in treatment and control schools who were eligible to receive the programme with similar pupils in control schools who were not. However, the randomization was undertaken at school level and hence does not guarantee that the characteristics of a subset of pupils in treatment and control schools are similar. In addition, a small number of schools dropped out of the programme, and not all pupils sat the ART post-test, which meant that the original randomization was compromised. Consequently, we had to try to 'balance' the characteristics of pupils in treatment and control schools, to ensure we were comparing like with like.

This was a challenging exercise, as the characteristics of pupils in treatment and control schools were very different. We identified pupils in treatment and control schools who were as similar as possible using a number of characteristics we deemed important for pupil progress such as prior attainment, gender, ethnicity, eligibility for free school meals, school type and composition.

Unfortunately it was not possible to find a group of pupils in control schools who 'looked' similar enough to pupils who were eligible to receive the LIT intervention across all of the dimensions that we thought mattered for progress. This meant that the original intention of estimating the impact of the LIT programme on the approximately 15% of Year 7 children in treatment schools who were eligible to receive it was not possible. Instead, the final analysis compared the reading scores of the 2,889 Year 7 pupils in 19 treatment schools – 660 of whom were eligible to receive the LIT programme and 2,229 of whom were not – with the 1,524 Year 7 pupils across 15 control schools who provided useable data at the start and end of the trial.

If we instead thought that only prior attainment (and specifically Key Stage 2 scores) mattered for pupil progress then we could estimate the impact of the LIT programme on pupils who were eligible to receive it, by comparing those pupils in treatment and control schools with similar Key Stage 2 scores.

Doing so resulted in an estimate of the impact of the LIT programme on those who are *eligible* to receive it (but not necessarily receiving it) of 0.13 standard deviations. However, this estimate is not statistically different from zero, and is highly likely to be biased by the fact that the pupils that we are comparing differ in other ways that matter for their progress. We would therefore strongly caution against regarding this as a robust estimate of the effectiveness of the LIT programme on those who received it.

Recommendations for further study

In addition to a further study to estimate the impact of the LIT programme on pupils who received it, it would be useful to explore in more detail which elements of the programme are vital to its success. The LIT programme was delivered in a variety of ways across schools in the evaluation, with variation in programme length, number of pupils per group, qualifications and experience of staff delivering the programme and whether they were specialists in English, and whether the LIT programme was delivered in addition to or instead of English or literacy lessons. Unfortunately it was not possible to separately identify the contributions of these different elements to the overall programme impact.

What did schools think of the programme?

The reciprocal teaching approach on which the LIT programme is based was seen as having a number of strengths: it encourages teamwork and debate, gives lower attaining pupils confidence, and is fun. But it was not felt to be suitable for the lowest ability pupils who required intensive support. Teachers felt that the positive benefits of using the LIT programme arose from the reciprocal teaching approach, the fact that teaching occurred in small specialised groups, the engaging programme content and quality of resources, and the assessment and pupil feedback element. Several schools reported that they would continue to use the programme at the end of the intervention.

How much does it cost?

The cost of training and detailed materials for the programme is £3000 per school. This price includes all of the planning and pupil resources, grammar and punctuation booklets for teachers and pupils, a handbook, on-site training for up to 20 members of staff, and follow-up support via email or phone. The other major costs are likely to be staff time: this will vary depending on whether a teacher or teaching assistant delivers the programme, whether the lessons are given in addition to or in place of English lessons, what the staff member would have been doing instead of teaching the LIT programme, and the number of pupils treated.

Introduction

Intervention

The LIT programme aims to help children in Year 7 who start secondary school with below average literacy skills (pupils with a national curriculum level of 2 or 3 in English at Key Stage 2, and in particular those whose reading has been assessed at Levels 2a, 2b or 3c at the start of Year 7). The programme aims to give students a chance to improve their comprehension skills, and provides staff with the training and support they need to deliver basic instruction in literacy. The programme – which can be delivered by teachers or teaching assistants, to whole classes or small groups – is founded on reciprocal teaching methods, and is generally used to supplement English lessons, with schools typically delivering 3-4 hours of the programme per week over approximately eight months.

Background evidence

LIT uses a mix of evidence-based approaches, including metacognitive methods and effective feedback,¹ but at its core is a method called reciprocal teaching. Reciprocal teaching is an instructional procedure designed to teach students cognitive strategies that might lead to improved reading comprehension. Discussion between teacher and students is used to help students learn strategies such as summarisation, question generation, clarification, and prediction.

There is some previous non-experimental evidence on reciprocal teaching: a meta-analysis of 16 studies (including some randomized trials) found an average effect size of 0.32 standard deviations on reading test performance, and 0.88 standard deviations on experimenter-developed comprehension tests (Rosenshine and Meister, 1994). These are large effects, but the underlying studies all assessed slightly different treatments, and were not of uniformly high quality. There is thus still some uncertainty over the precise impact of reciprocal teaching, which this evaluation sought to address.

Reciprocal teaching methods had also not been trialled extensively in the UK and, given the policy context –18% of children leave primary school without achieving Level 4 in reading – it is particularly important to identify effective literacy interventions suitable for children in early secondary school. It was therefore deemed appropriate to conduct an efficacy trial of the LIT programme.

Evaluation objectives

The primary objective of the independent evaluation was to test whether the LIT programme had a significant impact on the literacy skills of the 15% of Year 7 children who took part. It also aimed to explore the potential for ‘spillovers’: whether Year 7 children in treatment schools who did not receive the intervention are affected because other children in their class or year group have received it. And to explore whether there was any variation in the impact of the LIT programme on different children. This impact evaluation was complemented by a process evaluation, which aimed to provide a detailed understanding of how the LIT programme was implemented, to identify the elements that are critical to its success, and to highlight any potential challenges with any future rollout of the programme.

¹ The EEF Toolkit states: “Feedback studies tend to show very high effects on learning. However, it also has a very high range of effects and some studies show that feedback can have negative effects and make things worse. Studies reporting lower impact indicate that it is challenging to make feedback work in the classroom. Meta-cognitive and self-regulation approaches have consistently high levels of impact with meta-analyses reporting between seven and nine months’ additional progress on average. It is usually more effective in small groups so learners can support each other and make their thinking explicit through discussion.” Source: <http://educationendowmentfoundation.org.uk/toolkit/>

The key research questions to be addressed by the evaluation were therefore:

- What was the impact of the LIT programme on the literacy skills of children who received it?
- Did this vary across children, e.g. by demographic characteristics or prior attainment?
- Is there any evidence that other Year 7 children in schools in which some individuals received the LIT programme were also (positively or negatively) affected by it?
- Which aspects of the programme appeared to be critical for its effectiveness?
- What lessons can we learn from the way in which the programme was implemented in these schools for any potential future rollout?

Evaluation team

The evaluation team comprised researchers from the Institute for Fiscal Studies and NatCen Social Research. IFS researchers were responsible for the design, conduct, analysis and reporting of the results of the impact evaluation. Researchers from NatCen Social Research worked alongside the intervention team to manage communication with schools; they were also responsible for designing and overseeing the testing protocol at baseline and follow-up, as well as the process evaluation, and for reporting on the results of this aspect of the evaluation.

Lead researcher from IFS:

Claire Crawford, Research Fellow (previously Programme Director) at IFS

Supported by:

Michael Webb, Research Economist at IFS

Haroon Chowdry, previously Senior Research Economist at IFS (now Evidence Analyst at the Early Intervention Foundation)

Lead researcher from NatCen:

Amy Skipp, Research Director at NatCen

Supported by:

Meg Callanan, Senior Researcher, Children and Young People Team, NatCen

Sarah Haywood, Researcher, Children and Young People Team, NatCen

Intervention team

The intervention team was responsible for school recruitment, intervention development and intervention delivery. They also helped to maintain relations with school contacts, with whom they had built up a constructive working relationship.

Elina Lam, Learning Trust

Sophie Holdforth, Learning Trust

Ethical review

Ethical approval for this study was obtained from two sources:

- NatCen Social Research obtained ethical approval from their own ethics board regarding communication with schools and the opt-out process for the trial, as well as the testing strategy and process evaluation.

- IFS obtained ethical approval from the University College London ethics board for data sharing (among the project team and with DfE) and for the procedures through which consent was obtained to link data to the National Pupil Database.

Acknowledgements

The evaluation team would like to thank the EEF, the implementation team, all pupils, staff and schools involved in the programme, and the NPD team at the Department for Education for their invaluable assistance in enabling us to evaluate the LIT programme. The IFS team would also like to thank members of the Education and Skills Sector at IFS, in particular Dr Barbara Sianesi, for much helpful advice and support throughout our work on this project, as well as two anonymous peer reviewers for helpful comments and suggestions. NatCen are grateful for the support of the Children and Young People team, especially Dr Emily Tanner.

Methodology

Trial design

The evaluation was conducted using a randomized controlled trial (RCT). Randomization took place at the school level. The main reasons for randomizing at the school rather than individual level included:

- *Fear of contamination*: because the programme trains staff in a new method of teaching, it could potentially be difficult to prevent those methods from being applied to other Year 7 pupils as well.
- *The desire to have a long-term control group*: this is relevant because it was decided to use a wait-listed control group approach (see below for further discussion).

There was a single treatment arm, with half of the 41 recruited schools randomly allocated to the treatment group and half to the control group. The control schools were placed on a wait list to receive the treatment for their new Year 7 cohort in the year after the evaluation ended. In treatment schools, a subset of Year 7 pupils was selected to participate in the intervention. Both the size of this group and the method through which pupils were selected to participate varied by school: some used pupils' scores on a baseline test; others used Key Stage 2 scores; others used pre-existing 'nurture' groups. This is discussed further below. In principle, however, Year 7 pupils were chosen to participate if they were deemed to have poor literacy skills (typically defined as scoring below the expected level in reading at Key Stage 2). In this sense the trial was a 'pragmatic' trial, testing the impact of the programme as the delivery organisation intended it and usually delivers it.

It was decided that all Year 7 pupils in treatment and control schools would be tested at the end of the intervention period (the end of Year 7) in order to estimate any 'spillover' effects. These are effects on pupils in treated schools who did not receive the treatment themselves. In the event, this decision turned out to be quite important, as there was a degree of fluidity in the group of treated pupils in some schools, meaning that those selected at the beginning of the year were not always those who were still receiving the treatment at the end of the year.

Pupils were tested at baseline as well as follow-up, for two reasons: first, the intervention team believed that identifying changes in particular types of literacy skills compared to the relatively broad baseline measures available from Key Stage 2 tests would have been very challenging. Second, baseline testing was a means of identifying eligible children more accurately. The Hodder Access Reading Test (ART) was used. We discuss in more detail below some of the challenges with this test.

There were no significant changes to the design of the trial, but there were some issues with attrition. There were also problems ensuring that all pupils could be tested at both baseline and follow-up in a timely fashion. And, as mentioned above, there appears to have been some fluidity over which children received the LIT programme. All three aspects have consequences for the estimation and interpretation of the programme impact. We discuss these issues in more detail in the Conclusion.

Eligibility

School inclusion criteria: schools were recruited by the intervention team. The aim was to recruit 40 schools with high proportions of pupils eligible for free school meals and high proportions of pupils experiencing difficulties with literacy (or low proportions of pupils reaching Level 4 in reading at Key Stage 2). These aims were broadly achieved: on average, 34% of pupils in schools participating in the

LIT evaluation were eligible for free school meals (compared to 16% nationally) and 15% did not achieve Level 4 in reading at Key Stage 2 (compared to 14% nationally).²

Initially, all schools were going to be recruited from within London, but this was relaxed towards the end of the recruitment period, with a small number of participating schools from the South East and South West of England. A total of 41 schools signed up to participate in the trial.

Schools were required to sign a memorandum of understanding explaining the intervention, evaluation methodology and what would be expected from their involvement. Copies of these documents are included in Appendix 5.

Pupil inclusion criteria: discussions with the intervention team suggested that scoring in the bottom 25% of the national distribution of ART scores would be a sensible criterion to use to select pupils to participate in the LIT programme. However, the tight timescales involved in setting up the trial meant that a more pragmatic approach to selection had to be adopted. To try to maintain some uniformity, schools were asked to adopt one of three potential approaches³:

- **Option A:** include all pupils scoring Level 2 or 3 in Key Stage 2 reading (approximately equivalent to scoring below the 25th percentile of the ART test);
- **Option B:** if they had concerns about their ability to deliver the intervention to all pupils meeting the above criteria (e.g. because of capacity constraints), then they could instead offer it to a random selection of (at least 16) pupils scoring Level 2 or 3 in Key Stage 2 reading;
- **Option C:** some schools had already organised their timetables and would have found it difficult to accommodate the changes necessitated by either of the above choices, so they were additionally given the option of offering the intervention to a pre-selected 'nurture' group of children deemed to be in need of literacy support.

Schools were asked to state their choice of approach before randomization, so that this information could be used to help identify an equivalent group of potentially eligible pupils in control schools. Thirty schools reported that they would follow Option A, 4 schools Option B, and 7 schools Option C. This choice was taken into account in the randomization procedure. In practice, however, schools adopted a wider variety of approaches to recruitment. We discuss these issues further below.

Schools were asked to provide details of the pupils selected for the intervention at the start of the trial. The majority scored below the 25th percentile on the baseline ART test, although some scoring below this level did not receive it and some scoring above this level did. We discuss these issues in more detail below.

All parents of children in Year 7 were given the opportunity to opt out of the trial. We also sought permission to access the data held on their children by the Department for Education as part of its National Pupil Database. Consent was sought after randomization had taken place and no parents chose not to take part. See Appendix 5 for details.

Intervention

² Sources: <https://www.gov.uk/government/publications/schools-pupils-and-their-characteristics-january-2013>; <https://www.gov.uk/government/publications/national-curriculum-assessments-at-key-stage-2-in-england-2012-to-2013-provisional>.

³ It should be noted that these criteria are different from those previously used by the intervention team when selecting pupils to participate in the programme in other settings. This had previously been done using a combination of Key Stage 2 test scores and their performance on the Hodder Single Word Reading Test. It was deemed unfeasible to adopt this approach at scale given the tight timescales involved.

The LIT programme aims to help children in Year 7 who start secondary school with poor literacy skills. It offers training to staff, as well as a set of detailed lesson plans, on the use of reciprocal teaching, the core approach underlying the LIT programme.

The teacher or teaching assistant and pupils take turns to lead a discussion of a text adapted from children's literature, with the teacher showing children how to engage with the writing in increasingly sophisticated ways. Specifically, the teacher shows the children how to apply four comprehension strategies: summarising, clarifying, questioning, and predicting. These strategies force children to check that they understand the content, to identify exactly what the problem is when they don't understand, and to make inferences based on what they have read.

The method of delivery for the LIT programme is not particularly prescriptive: it can be used to teach whole classes or small groups, can be delivered by teachers or teaching assistants, and can be offered in addition to or instead of regular English classes (although children typically receive 3–4 hours of instruction as part of the LIT programme each week). This means that there is significant variation in the way in which treatment schools delivered the LIT programme as part of the trial. We discuss these issues in the process evaluation section below.

Schools which were randomized into the control group acted as a wait-listed 'business as usual' control group. This means that they were not prevented from undertaking other literacy interventions or changing their approach to literacy, just as they would not were the LIT programme to be offered more widely. The types of activities undertaken by pupils in control schools were investigated as part of the process evaluation, and are discussed in more detail in that section.

To encourage schools to join and continue participating in the evaluation, those allocated to the control group were offered the opportunity to receive training in the LIT programme at the end of the intervention at zero cost to them.

Outcomes

The primary outcome was the overall standardised score achieved on the Hodder Access Reading Test.⁴ This test was chosen because it was one of the few available on the market at the time that was deemed appropriate for capturing general reading ability and comprehension. Pupils sat the test either on a computer (29 schools) or on paper (5 schools), as some schools encountered difficulties using the digital version of the test.⁵ Schools were instructed that the tests should be taken under exam conditions, and the testing team undertook observations in a small number of schools to ensure that the testing protocol was being adhered to.

The tests were mostly blind marked. In the case of the computerised tests, standardised scores were automatically generated, downloaded by the school staff, and then sent on to the evaluation team. In the case of the paper tests, the marking of the follow-up test was undertaken by external assessors appointed by the evaluation team who did not know which schools or pupils were in which arm of the intervention. In the case of the baseline tests, some marking was undertaken by the intervention team, but these scores were verified by the evaluation team. We are confident about the marking of these scores in all cases.

⁴ For more details, see: <http://www.hoddertests.co.uk/tfsearch/ks2/reading/AccessReadingnew.htm>.

⁵ Only 34 of the original 41 schools provided baseline and follow-up test data for at least some pupils. We discuss the consequences of this attrition in more detail below.

Children participating in the LIT programme also undertook a single word reading test (Hodder Oral Reading Test) at both the start and end of the intervention.⁶ This test was carried out by the staff members delivering the programme within each school and hence was not blind marked. This test was not collected among a similar group of pupils in control schools, and hence does not play any part in the assessment of the effectiveness of the LIT programme; instead, it was intended to be used by the schools themselves to better understand the progress made by pupils receiving the intervention. Very few schools reported the results of this test to the evaluation team, so they are not analysed here.

Sample size

A target sample of 40 schools, to be split equally across treatment and control schools, had been agreed before the evaluation team joined the project. The power calculations were therefore conducted on this basis.

The published protocol (shown in Appendix 3) illustrated the sample sizes required for various combinations of the effect size and intra-cluster correlation, at intervals of 0.1. To illustrate the power required to identify smaller effect sizes, we expand upon these calculations below.

Table 1 shows the total sample size required (the number of children required in total across treatment and control schools) for 80% statistical power assuming there would be 20 treatment schools (clusters) and 20 control schools. Within the table, the required sample varies according to the estimated effect size (measured in standard deviations) and the within-school correlation in test scores. The table assumes a residual variance in follow-up test scores (that is, the unexplained variation in follow-up ART scores after controlling for baseline characteristics, including baseline ART scores) of 25%, and a cluster coefficient of variation (that is, an allowance for differences in school (cluster) size) of 0.3. (The published protocol assumed a cluster coefficient of variation of 0.25.)

Table 1 Sample size required for 80% statistical power

		Within-school correlation				
		0	0.05	0.1	0.15	0.2
Effect size (SDs)	0.05	3,140	N/A	N/A	N/A	N/A
	0.1	785	N/A	N/A	N/A	N/A
	0.15	349	632	6,354	N/A	N/A
	0.2	196	254	380	843	N/A
	0.25	126	144	172	219	318
	0.3	87	94	103	115	133

Calculations assume 20 each of treated and control schools. N/A means the required statistical power is not feasible given the effect size and within-cluster correlation because there are too few clusters.

The original intention was for the primary impact of the programme to be estimated on pupils who were eligible to receive it. Eligibility was supposed to be determined on the basis of scoring in the bottom 25% of the ART test, so the power calculations assumed that 25% of pupils in treatment and control schools would be used. With an average of 176 pupils in Year 7 across the schools who

⁶ For more details, see: <http://www.hoddertests.co.uk/tfsearch/ks2/reading/hort.htm>.

participated in the trial, Table 1 shows that, where detection at 80% power is feasible, there was a *priori* a good chance that an overall impact of the intervention on pupils in treatment schools could be detected.

Were the intra-cluster correlation – that is, the correlation in outcomes between pupils in the same school – to be greater than zero, however, it quickly becomes impossible to identify effect sizes at the smaller end of the spectrum, because there are too few clusters (see Appendix 3 for further details).

In the achieved sample of 34 schools, the intra-cluster correlation of follow-up ART test scores was 0.075 and the residual variance was 0.34. This means that there would be an 80% chance that an effect size of at least 0.16 standard deviations could be detected at the 5% level of significance, with a required sample size of 3,746.

As we shall see below, it was unfortunately not possible for us to estimate the impact of the LIT programme on the group of pupils were eligible to receive it in a robust fashion; instead we had to estimate the impact of the programme on all pupils in Year 7. This increases power, as it increases the number of pupils included in our estimates. However, it is important to bear in mind that only a small proportion of pupils in treatment schools actually received the LIT programme: around 15% of our eventual sample of Year 7 pupils.⁷ Assuming that the LIT programme had no effect on pupils in treatment schools who did not receive it, the impact on those pupils who did would need to be much larger than that cited in the power calculation tables to be detectable with the sample sizes at our disposal.

For example, to be able to detect an average effect size of 0.16 standard deviations, the effect on those receiving the intervention would have to be more than 1 SD (assuming that the programme had no effect on other pupils in treatment schools). If instead there were small positive spillover effects (i.e. small positive effects among pupils in treatment schools who did not receive the LIT programme – arising, for example, from the fact that their English classes might have been smaller than usual, or could go at a faster pace, because those with poorer literacy skills were doing LIT instead) then the necessary effect size among those receiving the intervention would be less than 1 SD.

Randomization

The randomization occurred at school level and was conducted using an iterative process in order to find an optimal randomization that ensured schools allocated to the treatment and control groups were as similar as possible in terms of their characteristics (e.g. school type and size) and the make-up of their student bodies.

The original sample of 41 schools participating in the trial was stratified according to region, local education authority (LEA), and LIT recruitment option (i.e. whether all pupils scoring below Level 4 in Key Stage 2 reading received the LIT programme, a subset of those pupils, or a pre-defined nurture group) before randomization.⁸ Where there was only one school within a region-LEA-option block, that school was assigned to the treatment group with 50% probability. Where the number of schools in that block was even, half were assigned the treatment group; where the number was odd, then with 50% probability schools were randomly assigned to the treatment group and with 50% probability they were randomly assigned to the control group.

⁷ These were the pupils reported by the schools to have been allocated to the LIT programme at the start of the trial. Given the movement of pupils into and out of the LIT programme throughout the year, this number should be taken as indicative only.

⁸ The sample was stratified on the basis of geography in order to account for any differences across local education authorities that might have affected children's literacy skills.

Random assignment guarantees that, on average, one should expect there to be small and statistically insignificant differences between the characteristics of treatment and control groups. However, this guarantee does not apply to any single random draw: in any particular draw it is possible that larger, significant differences can arise purely by chance. The evaluation team thus used an iterative procedure to identify an 'optimal' random assignment.

For each iteration, the process outlined above was carried out and two diagnostic checks were performed. First, the characteristics of treatment and control groups were compared and the number of statistically significant differences recorded.⁹ Second, the difference in average characteristics between the two groups was calculated.¹⁰ For each iteration, these two numbers were stored.

The iteration was carried out 1,000 times, resulting in 1,000 different treatment allocations. To identify the optimal randomization, the random assignments that led to zero significant differences between the two groups in terms of their characteristics were retained. Among this set of assignments, the one that had the smallest value of the total difference in mean characteristics was then selected. This was the final LIT treatment allocation that the evaluation team shared with the project team. The randomization was conducted blind by the evaluation team (i.e. we did not have access to school names when we were undertaking this process).

The full set of characteristics used as part of this process – together with the average differences in characteristics between treatment and control schools in our final allocation – are shown in Appendix Table A1. It should be noted that we only had access to school-level data at the time of randomization. This means that, in principle, if the characteristics of newly arriving Year 7 pupils are very different, on average, from the characteristics of pupils in the school as a whole, and in ways that differ across treatment and control groups, then it is possible that the characteristics of Year 7 pupils in treatment and control schools might be significantly different, even if the average school-level differences are not. This is indeed what happened in this evaluation, as we discuss further below.

Analysis

Analysis plan

Primary analysis

With a randomized control trial (RCT), a simple comparison of average test scores at the level of randomization (in this case school level) should, in principle, provide a valid estimate of the impact of the programme.

However, we chose a least squares regression at the pupil level, controlling for a variety of baseline characteristics, as the primary method of analysis in preference to this approach. This was because both the number of schools and the proportion of the variation in follow-up test scores that can be explained affect the power to detect significant effects. As the number of schools (clusters) included in the trial was already relatively small, these considerations seemed paramount. Including control variables as part of a regression analysis also minimises the risk that small differences in observable characteristics between the (randomized) samples affect the impact estimates. Regression estimates are thus typically more robust and more precise than a simple comparison of means.

⁹ This was assessed by conducting a t-test for each characteristic between the two groups of schools.

¹⁰ This was specified as the Euclidean norm of the vector of the standardised difference in mean characteristics between the two groups.

Moreover, our main population of interest was the relatively small group of pupils within each school who would receive the LIT programme. The primary analysis plan was thus to estimate the ‘intention to treat’ effect of the LIT programme by comparing pupils in treatment schools who were expected to receive the LIT programme with pupils in control schools who would have received the programme had they been in a treatment school. The plan was to identify these groups using ART baseline test score: specifically, those scoring in the bottom 25%. The outcomes of these pupils would be compared using a least squares regression approach controlling for baseline school and pupil characteristics.

The plan was to include ART baseline test scores, as well as pupils' performances in Key Stage 2 English (variable KS2_ENGFINE in the NPD) and Key Stage 1 Reading (KS1_READPOINTS). A range of other pupil- and school-level characteristics from the National Pupil Database were also going to be included. The inclusion of baseline measures of attainment as control variables is essentially equivalent to comparing the change in attainment among (or the progress made by) pupils in treatment schools with the change in attainment among (or the progress made by) pupils in control schools. This is known as a ‘difference in differences’ estimate of the impact of the programme.

It is a matter of judgement as to which characteristics should be included as controls. In this context, one should account for all of the characteristics that are believed to affect pupil progress – and that might plausibly differ between treatment and control groups – as it is the progress of pupils in the treatment group relative to the progress of pupils in the control group that is used to judge the impact of the LIT programme. Any differences in characteristics between pupils in treatment and control groups that matter for progress but that we do not account for in our model might potentially bias our estimates of the impact of the programme.

Standard errors were to be corrected for clustering at the school level.

Secondary analyses

Alternative populations

To investigate whether the LIT programme had any ‘spillover’ effects on Year 7 pupils in treatment schools who did not receive the programme themselves, we also planned to run the above regression analysis on all Year 7 pupils in treatment and control schools.

To investigate whether the impact of the programme varied across different types of pupils, we also planned to re-run the analysis among other subgroups of interest, notably those eligible for free school meals, as well as by gender, ethnicity and English as an additional language (as literacy skills are known to vary between boys and girls, and those with different language backgrounds).

Alternative ways to account for differences in baseline characteristics

In the event that ordinary least squares regression analysis was not sufficient to account fully for any differences in baseline characteristics between the treatment and control groups, it was also planned to conduct secondary analyses using propensity score matching (PSM) techniques. This essentially ‘re-weights’ pupils in the control group so that they ‘look like’ pupils in the treatment group, thus hopefully removing (or at least reducing) any differences in characteristics. PSM is more flexible than OLS for a variety of reasons, including the fact that it enables the analyst to identify and drop any pupils in the treatment group for whom a suitable control group cannot be found (we refer to these pupils as ‘off the common support’), rather than inappropriately extrapolating outside the area of common support, as OLS does. Matching also has the advantage that it automatically provides a series of diagnostic statistics, enabling judgements to be made about how well ‘balanced’ treatment and control groups are after the re-weighting process has been applied.

Post hoc analyses

Various problems with the trial, detailed below, meant that OLS and PSM alone would have produced biased estimates of the impact of the LIT programme. To account for attrition and missing data, as well as some significant individual-level differences between the treatment and control groups (not fully accounted for by the school-level aggregates used for randomization), the evaluation team used multiple imputation and multi-stage matching techniques to produce the final estimate of the treatment effect. These procedures are explained in more detail in the analysis section and in Appendix 2.

Even adopting these new approaches, however, we were unfortunately unable to ‘balance’ all of the characteristics of pupils in treatment and control groups who were eligible for the LIT programme (those scoring in the bottom 25% on the ART pre-test) that we judged to matter for progress (including gender, ethnicity, eligibility for free school meals, school type and composition). This means that we could not be sure we were comparing like with like on all these dimensions, thus potentially invalidating any estimates of the programme conducted on this group. We therefore departed from the evaluation protocol and focused on estimating the impact of the LIT programme on all Year 7 pupils as our primary analysis. These estimates will capture the average impact of the programme on the small group of pupils in treatment schools who received it, as well as the larger group of pupils who did not.

As outlined above, we believe that it is important to account for differences in a wide range of pupil and school characteristics that are likely to affect pupil progress and that may potentially differ between treatment and control schools. One could, however, argue that some characteristics are more important than others: for example, it might be reasonable to suppose that pupils’ baseline test scores are the most important factor in predicting how much progress they are likely to make over the coming year. If that was the case, then we could potentially set aside our concerns about the other ways in which pupils from treatment and control groups differ and focus on ensuring that they are as similar as possible in terms of baseline attainment.

With this in mind, we additionally produced estimates of the impact of the LIT programme on pupils who were eligible to receive it (those scoring at or below the 25th percentile on the baseline ART) by finding a group of pupils in control schools whose Key Stage 2 performance was as similar as possible to the eligible pupils’ in treatment schools. To the extent that these estimates ignore differences in other characteristics that matter for pupil progress, however, they risk producing biased estimates of the impact of the programme by not comparing like with like. We would therefore advise readers to treat these estimates with particular caution.

Process evaluation methodology

The aim of the process evaluation was to explore schools’ views and experiences of delivery and fidelity to the programme. It involved:

- A short web-based survey administered to all participating schools to gather information on how LIT was delivered;
- 15 follow-up depth interviews with treatment schools to gather in-depth data on school views on and experiences of LIT, including how closely schools adhered to the programme and the nature of any differences in the way it was delivered;
- 10 follow-up depth interviews with control schools to explore the nature of other literacy support provided to pupils that may have had a bearing on the impact analysis.

The interviews were based around a topic guide to ensure systematic coverage of key issues, but were also intended to be flexible and interactive, allowing issues of relevance for individual respondents to be covered through detailed follow-up questioning. The interviews were digitally recorded and subsequently analysed using Framework, a systematic approach to qualitative data management developed by NatCen Social Research and now widely used in social policy research.

All participants were told that everything discussed in the interview would remain confidential and would be treated in accordance with the Data Protection Act. Additionally it was made clear, both on recruitment materials and during the interview, that their views or opinions would not be discussed outside of the research team, including sharing individual feedback with the intervention team. All collection and analysis of data was conducted by the independent NatCen process evaluation team.

Impact evaluation

Timeline

School recruitment took place up to June 2012. Schools were randomized in July 2012. The original plan was for baseline testing to be carried out during a two-week period following the return from the summer holidays. However, schools had very little time to prepare for the baseline testing and some struggled with both timetabling and the technical requirements involved in delivering the digital tests. The testing period therefore had to be extended and some baseline test data was not received until the end of November 2012.

The intervention was due to run from September 2012 for a full academic year (i.e. until the end of July 2013), with the follow-up test administered in June 2013. Owing to the issues mentioned above, some participating schools did not complete their baseline testing until after they had started the LIT programme, and others started the programme late (up to two months after they were originally due to start). Learning on all sides from the baseline testing process meant that the follow-up testing period ran more smoothly, and the vast majority of schools completed and submitted their test data on time.

Recruitment and follow-up of participants

The intervention team recruited schools to participate in the trial. As described above, a total of 41 schools, mostly from within London and with higher than average proportions of pupils eligible for free school meals, were recruited. This was slightly above the target of 40 schools.

Attrition

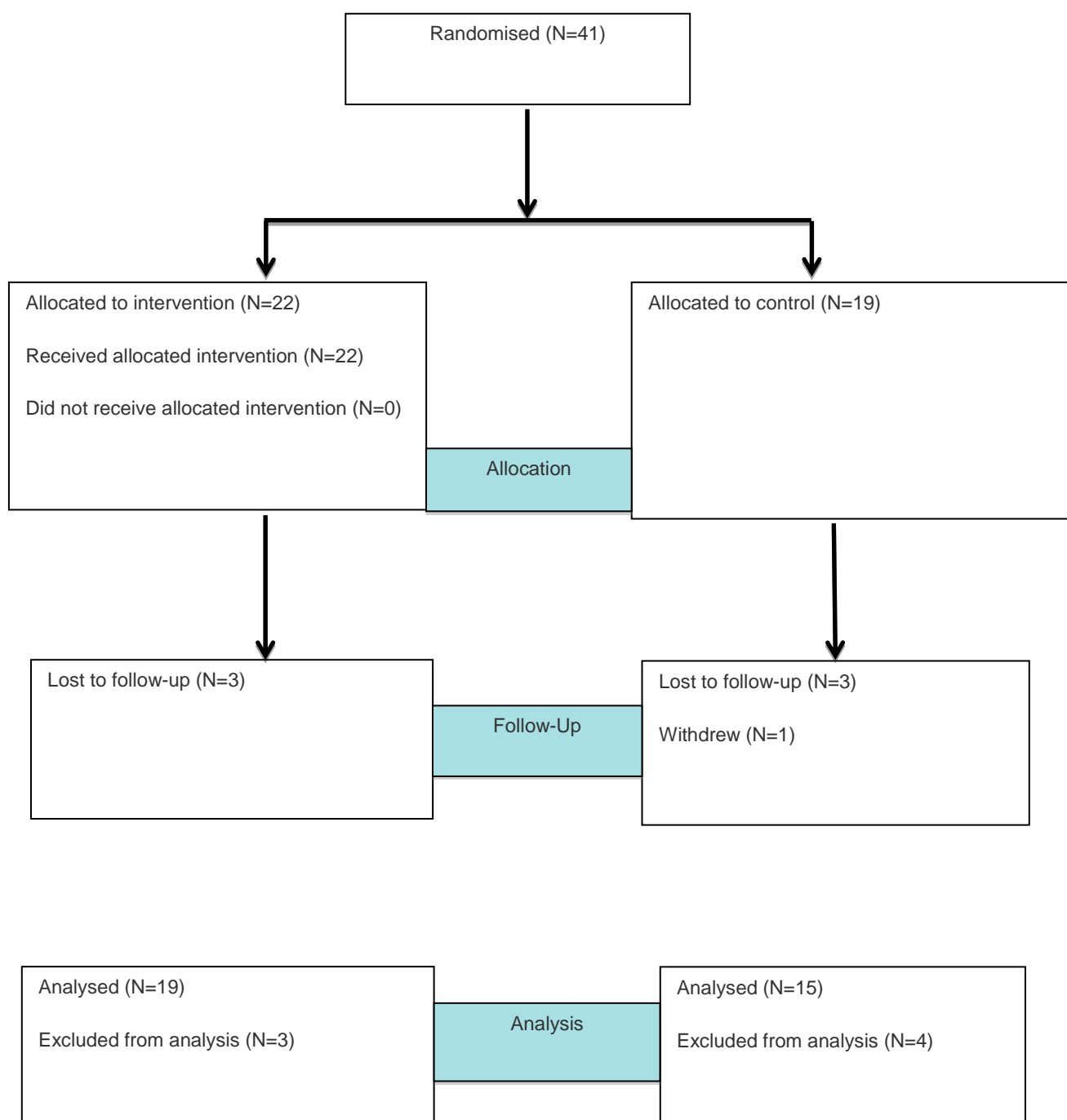
A number of schools dropped out of the intervention or else failed to provide either baseline or follow-up test data, which meant that they could not be included in the analysis. This applied to four of the 19 control schools and three of the 22 treatment schools. To our knowledge, only one school formally withdrew from the programme; the remainder had difficulties implementing the testing.

This process is described pictorially in Figure 1. This diagram only includes figures for school attrition, not pupil attrition, as this is the level at which randomization was conducted.

The schools that dropped out differed from those that did not. The characteristics of schools which did and did not drop out are reported in Table 2. In particular, it is clear that different school types had very different propensities to remain in the trial, with, for example, almost half (3/7) of sponsor-led academy schools dropping out but fewer than a fifth (3/16) of community schools doing so.

In addition, even among schools that remained in the programme, not all pupils sat both baseline and follow-up tests: 92% of pupils sat the baseline test, 86% sat the follow-up test and 79% sat both the baseline and follow-up tests. There were a number of reasons for this missing data, mainly associated with the challenges that schools faced in carrying out the testing, e.g. difficulty accessing the software, difficulty accessing the necessary space and resource to carry out large-scale testing, etc. While we would, a priori, not anticipate the characteristics of pupils who were and were not affected by these issues to differ systematically, when the scale of the challenge in securing usable test data became apparent, we advised struggling schools to focus on testing pupils that were likely to be eligible for the LIT programme, as this was the group on whom we were planning to estimate the impact of the programme. This will, of course, introduce systematic differences.

Figure 1 Intervention flow diagram



We summarise below the implications of these problems of attrition and lack of data for the differences in characteristics between the treatment and control group as analysed compared to at randomization.

Table 2 School characteristics by dropout status

Variable	Non-dropouts	Dropouts	Difference
Headcount	1098	994	-104
KS2 avg. point score	26.04	26.43	0.39
% 5 A*-C	55.1%	56.3%	1.2%
London	76.5%	71.4%	-5.0%
Academy converters	20.6%	0.0%	-20.6%
Academy sponsor-led	11.8%	42.9%	31.1%*
Community school	38.2%	42.9%	4.6%
Foundation school	14.7%	0.0%	-14.7%
Voluntary aided school	11.8%	14.3%	2.5%
Voluntary controlled	2.9%	0.0%	-2.9%
Mixed	79.4%	85.7%	6.3%
% FSM	30.2%	33.4%	3.2%
% EAL	36.9%	38.3%	1.4%
% White British	41.3%	35.1%	-6.2%

* indicates significance at 5% level, ** at 1%.

Targeting the programme

While schools had been asked to indicate how they would select pupils to participate in the LIT programme before the intervention started, many adopted different approaches in practice. These changes arose primarily for pragmatic reasons, such as challenges in conducting the baseline ART, or for timetabling reasons. In the event, schools ended up using one of four main approaches:

1. All pupils scoring below the 25th percentile on the Access Reading Test (ART) were selected for the programme.
2. A subset of pupils scoring below the 25th percentile on the baseline ART test were selected in schools where they were unable to accommodate all pupils scoring below this level. In some instances those selected to participate were the lowest scoring students. In other cases, other factors were used, including one or more of the following:
 - Teacher assessment of which students were struggling most with the mainstream curriculum;
 - Consideration of group dynamics (e.g. behaviour, gender mix);
 - Perceived ability to access the programme (e.g. not including children with limited English or those with Special Educational Needs).
3. Teacher assessment and Key Stage 2 results were used where schools were unable to use the ART results to select students for the LIT programme because the tests were conducted in the Autumn term and timetabling requirements meant groups had to be selected prior to this.
4. Some schools delivered the LIT programme to a pre-defined nurture group formed to meet a range of educational needs, including emotional and behavioural barriers to learning. In some cases this meant pupils scoring above the 25th percentile on the ART received the programme.

Teachers also appeared to vary the make-up of the group undergoing the intervention once the programme was under way. For example, while the LIT programme is aimed at pupils entering secondary school with lower than average levels of literacy, we found that teachers were deciding who to include based on how appropriate they felt the programme was on the basis of:

1. Literacy levels (some felt it was more suited to those not at the lower end of the target group);
2. Special educational needs (depending on the level of pupils' needs);
3. Those with English as an additional language (with those with less ability in English being felt to be unable to access the programme).

Baseline characteristics

School characteristics

Table 3 presents a selection of the characteristics of the 41 schools that were recruited to the trial ('At randomization') and the 34 schools that reported both baseline and follow-up test data for at least some pupils. A more detailed set of characteristics can be found in Tables A1 and A2 in Appendix A.

Table 3 School characteristics at randomization and as analysed

	At randomization			Schools with pre- and post-data		
	Control	Treated	Difference	Control	Treated	Difference
Headcount	1095.8	1067.5	-28.3	1124	1078.2	-45.8
KS2 avg. point score	26.8	25.5	-1.29	27.027	25.263	-1.76
% 5 A*-C	56.6%	54.1%	-2.5%	56.8%	53.7%	-3.1%
London	73.7%	77.3%	3.6%	73.3%	78.9%	5.6%
Academy converters	15.8%	18.2%	2.4%	20.0%	21.1%	1.1%
Academy sponsor-led	15.8%	18.2%	2.4%	6.7%	15.8%	9.1%
Community school	36.8%	40.9%	4.1%	33.3%	42.1%	8.8%
Foundation school	15.8%	9.1%	-6.7%	20.0%	10.5%	-9.5%
Voluntary aided school	15.8%	9.1%	-6.7%	20.0%	5.3%	-14.7%
Voluntary controlled	0.0%	4.5%	4.5%	0.0%	5.3%	5.3%
Mixed	84.2%	77.3%	-6.9%	86.7%	73.7%	-13.0%
% FSM	31.6%	30.0%	-1.6%	31.0%	29.7%	-1.3%
% EAL	36.0%	38.2%	2.2%	35.0%	38.4%	3.4%
% White British	40.8%	39.7%	-1.2%	44.0%	39.1%	-4.8%
Number of schools	19	22		15	19	

* indicates significance at 5% level, ** at 1%.

The table shows that the characteristics of the treatment and control schools were very similar at randomization. By definition, none of the differences at randomization were statistically significant. However, this does not ensure that the schools are similar in terms of other characteristics that were not available or observable to us. It is also clear that the differences are, in almost all cases, larger among the schools for which analysis can be carried out than among all schools at randomization. This stems directly from the fact that schools did not drop out randomly (as discussed above).

Pupil characteristics

There were 3,457 Year 7 pupils in total in the remaining 19 treatment schools and 2,108 in the remaining 15 control schools. A further 1,152 pupils, split approximately equally between treatment and control schools, could not be included in the analysis due to missing test scores. This gave a total sample size for the analysis of 2,889 pupils in treatment schools and 1,524 pupils in control schools, representing 79% of the total number of Year 7 pupils in those schools.

Table 4 presents a selection of the characteristics of Year 7 pupils in treatment and control schools at baseline. The table displays the data for all pupils in the 34 schools that reported (some) baseline and follow-up test data, and for all pupils with both individual scores (i.e. the analysis sample).

It is worth noting here that in addition to the differences in *school* characteristics between treatment and control schools increasing following school drop-out (see Table 3 above), there were also significant differences in *pupil* characteristics between the remaining treatment and control schools.¹¹ For example, there were significant differences in terms of the proportion of pupils eligible for free school meals and the proportion of pupils from some ethnic groups. It will be vital to account appropriately for these differences in order to produce an unbiased estimate of the impact of the LIT programme. We discuss these issues in more detail below.

Table 4 Pupil characteristics at randomization and as analysed

	All pupils in 34 reporting schools			All pupils with pre- and post-test scores		
	Control schools	Treated schools	Difference	Control schools	Treated schools	Difference
Baseline ART score	-0.03	-0.01	0.02	-0.02	0.02	0.04
KS2 English	4.63	4.58	-0.05*	4.62	4.61	-0.02
KS1 Reading	14.90	14.78	-0.12	14.86	14.94	0.09
Female	46.39%	48.21%	1.82%	45.89%	48.76%	2.87%
EAL	45.00%	46.49%	1.49%	42.32%	45.34%	3.02%
FSM	36.93%	32.31%	-4.62%**	35.65%	31.31%	-4.34%**
SEN	27.53%	30.14%	2.61%*	28.03%	28.91%	0.87%
White	50.50%	48.63%	-1.87%	52.03%	48.63%	-3.40%*
Asian	21.19%	15.66%	-5.53%**	17.00%	15.20%	-1.80%
Black	17.08%	17.59%	0.51%	16.86%	16.68%	-0.18%
Chinese	0.30%	0.60%	0.31%	0.33%	0.59%	0.26%
Other	10.94%	17.53%	6.59%**	13.78%	18.90%	5.12%**
Number of pupils	2,108	3,457		1,524	2,889	

* indicates significance at 5% level, ** at 1%.

It was not possible to present the baseline pupil level characteristics for the 41 schools randomized because the schools that dropped out did not provide consent to access their pupils' UPNs and National Pupil Database records. Therefore we are unable to analyse to what extent these differences were present at randomization, or occurred due to attrition.

As well as estimating the impact of the LIT programme on all Year 7 pupils in the school, it was also hoped that we would be able to estimate the impact on pupils eligible to receive the programme (broadly classified as those scoring below the 25th percentile on the baseline ART test). Selected characteristics of pupils scoring below the 25th percentile on the baseline ART test are displayed in Table 5. While there are only very small differences between pupils scoring in the bottom 25% in treatment and control schools in terms of the baseline ART test, there are larger, significant, differences in terms of other characteristics, such as key stage scores, free school meals eligibility and ethnic group.

There are also large and significant differences in combinations of characteristics among this group of pupils, which is what is relevant when it comes to 'balancing' the characteristics of pupils in treatment and control schools. For example, 47% of Asians in control schools are eligible for free school meals, compared with only 28% of Asians in treatment schools. Similarly, while 44% of treatment school

¹¹ Unfortunately, NPD data was not acquired for schools that dropped out.

pupils and 34% of control school pupils are eligible for free school meals in Academies, the situation is reversed for Foundation schools, in which 44% of treatment school pupils but 57% of control school pupils are eligible. These differences in combinations of characteristics are problematic because they make it difficult to find pupils in control schools who are similar *in all dimensions* – or at least on the basis of all the observable characteristics at our disposal – to pupils in treatment schools who were eligible to receive the LIT programme. Unfortunately, this ultimately meant that it was impossible to produce an unbiased estimate of the impact of the LIT programme on pupils scoring in the bottom 25% of the ART at baseline. We discuss the reasons for this in more detail below.

Table 5 Characteristics of pupils scoring below 25th percentile on baseline ART test

	All pupils with pre- and post-test scores			Pupils scoring below the 25 th percentile on the pre-test		
	Control schools	Treated schools	Difference	Control schools	Treated schools	Difference
Baseline ART score	-0.02	0.02	0.04	-1.36	-1.38	-0.02
KS2 Eng	4.62	4.61	-0.02	3.83	3.71	-0.12*
KS1 Reading	14.86	14.94	0.09	11.28	10.66	-0.62*
Female	45.89%	48.76%	2.87%	40.06%	43.45%	3.40%
EAL	42.32%	45.34%	3.02%	43.65%	51.92%	8.27%*
FSM	35.65%	31.31%	-4.34%**	51.11%	43.77%	-7.34%*
SEN	28.03%	28.91%	0.87%	58.84%	62.46%	3.62%
White	52.03%	48.63%	-3.40%*	56.73%	48.03%	-8.70%**
Asian	17.00%	15.20%	-1.80%	17.94%	13.64%	-4.31%
Black	16.86%	16.68%	-0.18%	14.25%	15.76%	1.51%
Chinese	0.33%	0.59%	0.26%	0.53%	0.30%	-0.23%
Other	13.78%	18.90%	5.12%**	10.55%	22.27%	11.72%**
% Asians with FSM	33.21%	22.55%	-10.65%**	47.06%	27.78%	-19.28%*
% Academy pupils with FSM	21.87%	31.81%	9.94%**	33.74%	43.52%	9.78%
% Foundation pupils with FSM	46.55%	38.10%	-8.46%	57.14%	44.44%	-12.70%
Number of pupils	1,524	2,889	N/A	379	660	N/A

* indicates significance at 5% level, ** at 1%.

Accounting for school dropout and missing data

Schools generally dropped out of the intervention – and pupils generally did not sit one or both of the baseline and follow-up tests – because of challenges in implementing the tests. The reasons why schools found this particularly problematic – such as poor IT infrastructure or staff skills – may well be correlated with children’s progress in reading. If we do not properly account for this attrition, then any analysis will obtain biased estimates of the impact of the intervention. In particular, bias can occur when the reasons for attrition differ between treatment and control schools (e.g. if less motivated control schools don’t administer the follow-up test whereas similarly unmotivated treatment schools do), or when they are correlated with the treatment effect (e.g. treatment and control schools don’t test the lowest ability pupils, when it is precisely these pupils for whom the impact is greatest).

As has been described, a number of schools dropped out of the trial completely, and baseline and follow-up test data was not available for all pupils. This means that the impact of the LIT programme can be estimated only on pupils who were in schools that remained in the trial and who sat both the baseline and follow-up tests. It should therefore not be interpreted as the average treatment effect

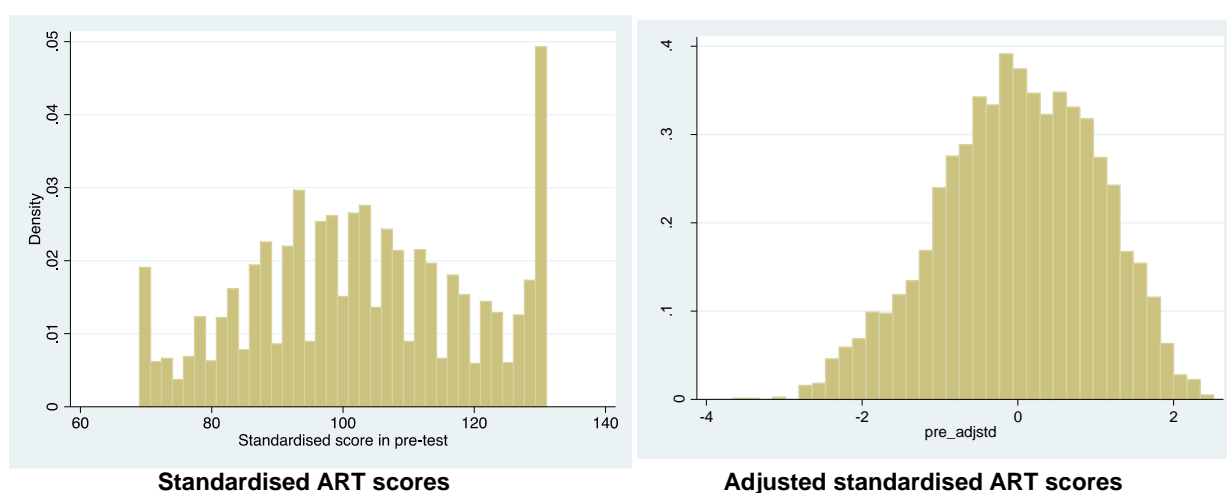
among the full sample.¹² The fact that the attrition of schools and pupils has generated significant differences in some observable characteristics between treatment and control schools means that it is even more important not to rely on a simple comparison of means between the two groups in order to estimate the treatment effect. This is because the underlying difference in characteristics will bias the resultant estimates. Instead, as outlined in the analysis plan, account must be taken of these differences in characteristics in some way.

In addition, there were 563 pupils for whom both baseline and follow-up test data was available, but additional information from the National Pupil Database (NPD) was not. For example, 437 of these pupils were missing Key Stage 1 scores and 97 Key Stage 2 scores.¹³ To avoid reducing our sample and biasing our results further, we imputed the missing data using a technique known as multiple imputation. This technique makes use of data from other pupils to account for the uncertainty arising from missing values. Figure A1 in Appendix 1 illustrates the extent to which our results would have been further biased by ignoring pupils with missing data, by showing that the distribution of Key Stage 2 test scores is likely to be very different among those with missing data compared to those for whom we observe this information. Multiple imputation is explained in more detail in Appendix 2.

Outcomes

As described above, the plan was to use the standardised score achieved on the Hodder Access Reading Test as the primary outcome measure of the evaluation. However, we encountered a number of problems with this measure. For example, the left-hand panel of Figure 2 plots the distribution of standardised ART scores. It shows that a very high proportion of pupils appeared to score at the ceiling and floor of the scale: 3.2% got the lowest score, and 8.3% the highest, against expected proportions of 2% in each case.

Figure 2 Distribution of ART standardised scores



¹² There were 397 individuals for whom we observed follow-up test data but not baseline test data. Because we could not be sure that these individuals were in the schools for all or part of the intervention period, we excluded them from our analysis.

¹³ Appendix Table A3 documents the patterns of missing data in more detail. It was deemed important to account for Key Stage 1 scores in addition to Key Stage 2 scores, because we only had access to information on Key Stage 2 levels (which are relatively coarse), so children awarded the same Key Stage 2 level might have been of very different underlying 'ability'. We additionally controlled for Key Stage 1 to try to overcome this potential bias.

This was driven in part by the way raw scores had been adjusted to account for variation in the age at which pupils sat the tests. As we had access to the raw scores and the age at which pupils sat the tests, we decided to recreate these measures 'from scratch', without imposing ceiling and floor scores. This was done by regressing raw scores on pupils' chronological age at time of test and using the residuals (the remaining variation unexplained by age at test) – standardised to have mean zero and standard deviation one – as our primary outcome measure.¹⁴ The resultant distribution of scores is displayed in the right-hand panel of Figure 2.

Estimates of treatment effect

As described above, we faced two key challenges in attempting to estimate an unbiased effect of the LIT programme on pupils' literacy scores:

1. Schools and pupils dropped out (or did not produce data) for a variety of non-random reasons.
2. As well as average differences in characteristics between the treatment and control group, we also found evidence of average differences in *combinations* of characteristics. This made it particularly difficult to 'balance' the characteristics of pupils in treatment and control groups.

For the planned analytical approach of ordinary least squares (OLS) regression analysis to be appropriate, it would need to be the case that controlling linearly for the ways in which individuals in the treatment group differed from individuals in the control group was sufficient to eliminate the differences. The 'propensity score' is a convenient way of summarising how 'different' individuals in the treatment and control groups are, as it provides an indication of how likely an individual is to be treated on the basis of their observable characteristics.

With a perfect randomized control trial (RCT), the average propensity scores in the treatment group would be the same as those in the control group: it would be difficult to tell just by looking at their characteristics whether a given individual was in the treatment or control group. In the absence of the perfect experiment, however, the average difference between the propensity scores in the treatment and control groups provides a good indication of whether OLS regression analysis is likely to control reliably for the differences and thus whether it is appropriate to use to estimate the treatment effect.

To more formally assess the balance between the treatment and control groups, we used two diagnostic measures. The first, B , is the number of standard deviations between the means of the propensity scores in the treated and control groups; as described above, this would be zero in an ideal RCT: the scores should be exactly balanced. The second, R , is the ratio of the variances of the propensity scores between the treatment and control groups. In an ideal trial, the variances of the two groups would be the same, so their ratio would be one.

Rubin (2007) suggests that treatment and control groups with a B score of greater than 0.3 are "too far apart to rely on linear regression adjustment", i.e. we should not rely on OLS when we find a B score greater than 0.3. The B score of this trial's raw data is in fact greater than one; its R score is approximately 0.3. The figures are similarly bad if we just focus on pupils scoring below the 25th

¹⁴ This procedure was undertaken separately for baseline and follow-up test scores; for the follow-up test, only pupils in control schools were used to fit the main regression, but residuals were obtained by predicting the outcome variable for all pupils. These age-adjusted scores were then standardised, this time pooling all baseline and follow-up test observations in both treated and control schools together. The distributions of these scores at baseline and follow-up test in treated and control schools is displayed in Figure A2 in Appendix 1. We did not simply control for age in our analysis, because the way the scores are calculated makes it vital to balance properly on age, and simply including it as one characteristic among many does not ensure that this will happen.

percentile in the baseline ART test (i.e. those who should have received the LIT programme), with a *B* score again greater than 1 and an *R* score of 0.2.

The figures are all very far from what we would hope for with an ideal RCT and imply that relying on OLS to estimate the impact of the LIT programme, especially for those in the bottom 25%, would almost certainly result in a severely biased estimate. Imposing common support (i.e. dropping pupils in the treatment group for whom no similar pupil can be found in the control group) and ‘re-weighting’ individuals in the control group so that they ‘look’ as similar as possible (in terms of the propensity score) to individuals in the treatment group improves but does not solve this problem in our case: we are still left with a *B* score of 0.6 and an *R* score of 2 for the sample overall and a *B* score of 1.2 and an *R* score of 1.4 for the bottom 25%.¹⁵ Again, using OLS would likely result in a biased estimate.

To deal with this problem, we used a procedure called ‘stratification on the propensity score’ (Rubin, 2007). This means that we compared the outcomes of smaller subgroups of our sample with more similar propensity scores (and hence which should, in principle, be more similar to one another). We assessed the *B* and *R* scores for each of these subgroups (referred to as ‘bins’) separately and then averaged them over all ‘bins’ to obtain overall *B* and *R* scores. Appendix 2 presents the method in more detail.

This procedure does a good job of balancing pupils in treatment and control schools when we use the whole sample (but not the bottom 25%: see below). Whereas the original *B* score for the whole sample was greater than 1 and the original *R* score around 0.3, the final *B* score is 0.1 and the final *R* score is 0.8. We are therefore confident that we can calculate an unbiased estimate of the impact of the LIT programme on all pupils in Year 7 in treatment schools. The box below outlines our key findings.¹⁶

Key findings:

- Our analysis suggests that the ART scores of Year 7 pupils in schools operating the LIT programme rose **0.09 standard deviations** faster over the course of the year than those of similar Year 7 pupils in control schools. This is equivalent to **1 month’s progress in reading**.
- But statistical tests indicate that **this estimate is not ‘significant’**, i.e. we cannot be sure that the difference in scores is not due to sampling error. We thus cannot conclude with certainty that the LIT programme had a positive effect.
- This estimate is **based on the outcomes of all pupils in treatment and control schools** (not just those pupils scoring in the bottom 25% on the baseline ART). It is therefore perhaps unsurprising that the effect size is smaller than previous estimates of the impact of reciprocal teaching (although this is not the only reason why the estimates might differ).

This effect size is much smaller than the average of those reported in a meta-analysis of estimates of the impact of reciprocal teaching (Rosenshine and Meister, 1994). There are a number of possible reasons for this. It is highly likely that it is at least partly driven by the fact that our estimate is an average of the effect among a relatively small group of pupils who actually received the LIT programme (in treatment schools) and a much larger group of pupils who did not (also in treatment schools). In this case, to the extent that the intervention had a larger positive effect on those who

¹⁵ This approach is known as propensity score matching. There are various ways in which it can be implemented (see the technical appendix for further details).

¹⁶ These estimates include baseline ART scores as control variables. If we were to omit these measures from our analysis, we obtain a larger and significant estimate of the impact of the programme. There are a number of reasons why this might be the case, and we discuss the relevant issues in more detail in Appendix 2.

actually received the programme, this might represent an underestimate of its effect on those pupils. On the other hand, it could be that the LIT programme was no better than the literacy support it replaced; that it was not implemented for long enough or with high enough fidelity to have a significant impact; or that there was too much transition into and out of the treatment group for the effects to be sustained (as described above). Our ‘whole cohort’ estimate has the advantage of capturing the effect of any such changes in treatment allocation.

There are also plausible reasons why our estimates might be overestimates of the impact of the LIT programme on children’s literacy skills, however. For example, schools in treatment areas might have been more motivated to take the testing procedure seriously, because they were more interested in knowing whether the LIT programme had been effective in raising achievement. If that meant that pupils in treatment schools (were pushed to) try harder in the follow-up tests, then this might upwardly bias our estimates of the impact of the LIT programme.

Unfortunately, it was not possible to produce an unbiased estimate of the impact of the LIT programme on those pupils who should have received it (i.e. it was not possible to conduct the primary analysis as originally specified in the evaluation protocol). This was because the characteristics of pupils scoring in the bottom 25% of the baseline ART in treatment schools were too different from those scoring in the bottom 25% in control schools to be confident that comparing their outcomes at the end of Year 7 would produce a sensible estimate of the impact of the programme on this group.

Nor was it possible to achieve sufficiently good balancing on all characteristics for other subgroups for which it was intended to produce separate estimates of the impact of the programme, such as those eligible for free school meals. Even the best specifications for these groups had *B* scores far beyond the threshold of acceptability. This means that the characteristics of individuals in these groups were too different for a comparison of their outcomes to generate a reliable estimate of the impact of the LIT programme on literacy scores, even using the advanced matching methods outlined above. Appendix 2 explains these issues in more detail.

As outlined above, however, deciding which characteristics matter for pupil progress – and hence must be balanced between treatment and control groups – is a matter of judgement. If we were to decide instead that pupils’ baseline test scores are the only factor that matters for pupil progress, then we could simply ensure that pupils in treatment and control groups were as similar as possible in this respect. The box below summarises the results of this exercise.

Secondary analysis:

- It was not possible to find a group of pupils in control schools who were similar in all respects that we believe matter for progress to pupils in treatment schools who were eligible to receive the LIT programme.
- If we thought that Key Stage 2 scores were the only factor that mattered for pupil progress, then we could estimate the impact of the LIT programme by comparing pupils with similar Key Stage 2 scores. Doing so resulted in estimate of the impact of the LIT programme on those who are *eligible* to receive it (but not necessarily receiving it) of **0.13 standard deviations**. However, **this estimate is not ‘statistically significant’**.
- Moreover, this approach ignores significant differences in other characteristics between treatment and control groups. To the extent that these characteristics matter for pupil progress, this will be a **biased estimate of the impact of the LIT programme** and should be treated with caution.

Cost

Detailed information on costs was not collected as part of the evaluation, as the myriad different ways in which the programme was delivered would have made an overall estimate of average costs

misleading. Some of the potential costs which should be borne in mind by schools when considering whether to take up the programme include:

- **Staffing and group size:** the small group nature of the teaching and the staffing costs associated with this were identified as one of the primary costs of the programme. Intervention schools adopted a number of approaches to delivering the intervention, including using only fully qualified teachers, only teaching assistants, or a combination of staff. These different options clearly have different cost implications and unfortunately it has not been possible to identify which may be the most cost effective.
- **Whether the programme replaces English lessons:** schools varied in how they chose to run the LIT programme: either replacing English lessons, or running it in addition to them. These alternatives would again have different cost implications, with the additional staffing requirements lower in schools running the programme as an alternative to mainstream English.
- **Training:** the initial requirement is for staff delivering the programme to attend a day-long training course. The cost of doing so (including materials) was £3,000 per school. This price includes all of the planning and pupil resources, grammar and punctuation booklets for teachers and pupils, a handbook, on-site training for up to 20 members of staff, and follow-up support via email or phone.
- **Resources:** the cost of purchasing resources was estimated to be £7 per pupil.
- **Administration:** costs associated with administering the programme, including identifying pupils, and planning the resourcing and timetabling of the programme.
- **Ongoing support:** additional consultations and support provided by the Learning Trust were estimated to cost £500 per time. Staff time may also be required to provide on-going support in order to deliver the intervention effectively.

Process evaluation

From the in-depth work carried out with participating schools, this section highlights key issues in implementation, delivery, teachers' views, and perception of impact of the LIT programme.

Implementation

Staff training and support

Prior to the start of the programme, training on the LIT programme was delivered by the Programme Manager (from the Learning Trust) to all schools delivering the intervention. The training typically lasted a day and provided an overview of the aims and objectives of the programme, the programme content, the principles of reciprocal teaching and the practicalities of implementation. The training was well received by staff who found it to be comprehensive and valuable.

Where suggestions for improvements to the training were made these included spending more time on the reciprocal teaching approach (as this approach was new to staff teaching the programme) and more time on the detailed content of each unit of work. Staff also suggested that a 'top-up' training session would be beneficial a few weeks into delivery to enable staff to share experiences and raise queries from their own teaching experience.

In addition to the initial training, the Programme Manager provided on-going support to schools in the form of email and telephone advice and site visits (including prompts on when and how to carry out testing for the evaluation). These visits also included lesson observations and advice and guidance on all elements of the programme. Again, this on-going support was valued by staff delivering the programme, particularly written feedback on lesson observations.

The resources provided for the LIT programme were viewed as comprehensive and compared favourably with other schemes of work in the level of detail:

"We were all given the LIT programme teacher's folder.. and I found that it was really, really useful. It's very thorough, which was really interesting because sometimes it can just be a bit vague, whereas I felt that there was lots of information, which was great."
(Newly Qualified Teacher)

A final suggestion for improving the support and training for the programme was to foster partnerships between schools delivering the programme to enable staff to share learning and observe LIT programme lessons delivered in different contexts by other teachers.

In summary, suggested improvements to the training included:

- Trainers spending more time on the reciprocal teaching approach
- Trainers explaining in more detail the content of each unit of work
- Trainers fostering partnerships between schools which would allow the sharing of learning and observations of lessons run in different schools.

Start dates

There was wide variation in the start dates for the LIT programme across the 19 intervention schools, with start dates typically ranging from September to November 2012. In rare instances, the programme did not start until the Spring term. Two reasons were given for delays:

- Delays in conducting the Access Reading Test: a lesson for future use of large-scale testing of this kind is the requirement for a planning and implementation phase to minimise delays of this kind.
- Staffing the LIT programme: where schools had staffing issues and found resourcing the LIT programme challenging, this delayed the start of the programme.

This meant that the length of time pupils participated in the programme varied by school (as they all finished at the end of the Summer term).

Intervention group make-up

Teachers were making decisions on who to include in their intervention group based on several factors. Sometimes these changes were made before the programme began and in other schools the changes were made during the programme.

Differing literacy levels

Teachers generally agreed that the programme was more appropriate for pupils with higher literacy levels than those at the lower end of the target group. In particular, they felt the programme worked best with pupils who were verbally articulate but had poor reading skills:

“One of the kids made six years’ progress in his reading comprehension age from being a non-reader to someone who’s picking up ‘Percy Jackson and the Lightning Thief’ for pleasure; because verbally he was really good but he just struggles with text.”

Teachers felt that for pupils who were initially on the borderline between level 3 and 4 the programme worked well to improve their reading and comprehension skills and to ensure that they were consistently performing at a level 4.

There were mixed views from teachers about how appropriate the LIT programme was for pupils working at lower levels, with conflicting views expressed on how well the programme worked for these pupils in terms of the difficulty level of the texts supplied by the programme and the interactive nature of the programme.

Pupils with Special Educational Needs

Views on whether the programme worked well with pupils with special educational needs (SEN) varied depending on the type and severity of the pupils’ need. Teachers identified the ability to communicate sufficiently to be able to contribute to the reciprocal teaching approach as vital to the success of the programme. For instance, different levels of autism in pupils was felt to yield different outcomes on the programme. For those who were low on the autistic spectrum teachers felt that the social interaction within the programme was beneficial and that they had made progress, but for those who were higher on the scale the programme was not suitable as they struggled not only with the more intense communication within the classroom but also with the abstract concepts explored within the texts.

Pupils with English as an additional language

Similarly, the success of the programme with pupils for whom English was an additional language (EAL) depended heavily on their knowledge of English and their ability to access the texts. For those who had no underlying cognitive issues but required more intensive support with vocabulary, teachers felt that the programme worked well:

“They’ve really flown because a lot of them are quite able and there’s no cognitive issues, they just needed a really systematic boost to their vocabulary and to cover all the bases.”

However, teachers felt that if pupils’ levels of reading and writing were too low to start with (e.g. pupils who had been in the country for less than two years), they failed to show progress on the programme as their low levels of reading prohibited them from accessing the texts at all.

There were three pupil ‘types’ for whom teachers felt the programme was most helpful:

- Pupils who are on the borderline between levels 3 and 4
- SEN pupils able to communicate enough to engage with the reciprocal teaching approach
- Pupils without underlying cognitive issues, requiring intensive vocabulary support.

Adaptations to the programme

Teachers were generally aware that the programme could not be adapted too much as it was being evaluated. However, they felt that if the programme was not working effectively with their pupils they wanted to make adaptations in order to ensure that their pupils did make progress.

There were four main ways in which teachers adapted the LIT programme. These adaptations were done to varying degrees in some of the intervention schools.

- **Homework:** teachers added homework to their lesson plans as they felt that homework was expected to be set for this subject and that doing the homework helped pupils to consolidate their learning from class. They also felt it was a helpful tool through which they could address individual skill weaknesses in pupils.
- **Basic writing skills:** teachers felt it was important to ensure pupils were working on their basic writing skills while they were on the programme. Some teachers therefore added exercises and additional work to their lesson plans as well as to their homework tasks.
- **Pace:** for a number of reasons some teachers needed to vary the pace of the programme. Generally this was because they either felt there was too much for them to cover within the time or that by adhering to the programme pace they would risk not meeting the programme objectives.
- **Varying the lesson content:** some teachers felt that the repetition and strict consistency of the lessons (particularly in terms of the reciprocal teaching approach) meant that pupils became bored or disengaged. Some teachers therefore altered the lesson plans, adding in additional exercises or not following the approach so closely in each lesson.

Delivery of the LIT programme

This section reflects on the variation in how schools delivered the LIT programme, with particular focus on staffing; group size and formulation; timetabling and frequency of lessons. Again these differences may account for some variation in effect, although due to small numbers it is not possible to identify any clear patterns.

Staffing the LIT programme

Across the schools delivering the LIT programme there was wide variation in how it was staffed:

- **Staff qualifications:** qualification levels of staff delivering the programme ranged from teachers with several years teaching experience to trainee teachers and teaching assistants. In some schools, only qualified teachers delivered the programme, while in others the programme was staffed entirely by teaching assistants. The choice of staff to deliver the programme was determined by a number of factors, with timetabling and staff availability key considerations. Some schools fed back that they felt the programme should be delivered by qualified teachers to provide the pupils with the most effective teaching. In other cases, it was felt teaching assistants with experience of delivering literacy support were able to deliver the programme effectively.
- **English specialism:** whether staff delivering the programme had an English specialism also varied. Because of timetabling and resources, there were some instances where staff delivering the programme were drawn from other specialities including Physical Education and Drama. In other instances, staff were selected because of their specialism in supporting children with SEN or EAL. Again, timetable availability and which school budgets were used to fund the staffing of the programme played a part in which staff delivered it. For less experienced staff and those without a background in teaching English, the set structure of the LIT programme and support from the Programme Manager was particularly valued:

“That's probably been better for us because of the level of experience or in some cases inexperience of some of the people doing it or the lack of familiarity with the English curriculum. I mean whilst I'm a qualified teacher I've never taught English before so it's been quite good to have that, that kind of framework and the support that's checking how we're doing.”

(Special Needs Coordinator)

Group size

Group sizes ranged from four to nine pupils, with some examples of larger groups being subdivided (for example a group of sixteen split into two groups of eight and supported by two members of staff). Schools took into account a range of factors when determining the composition of the groups:

- **Timetabling:** the practicalities of timetabling LIT lessons and the availability of classrooms and teachers was a determining factor in the composition of LIT groups in some schools.
- **Ability:** schools that grouped by ability did so to ensure the teaching and pace of the programme could be tailored to the needs of different ability groups. Other schools took the view that mixed ability groups would ensure the weaker students were supported by the stronger ones.
- **Group dynamics and behaviour:** staff felt it was important to pay attention to group dynamics, and factors taken into consideration included the gender mix, friendship groups and behaviour. Some schools also swapped groups around in cases where the group dynamics were not working, or where they felt the pupils would benefit from working with different personalities.

Groups of 5–6 pupils were felt to be the optimum size for effective delivery for the following reasons:

- The reciprocal teaching approach allocates five roles to pupils – clarifier, predictor, questioner, summariser and discussion director, so a group of this size works well with the approach
- Small groups allowed more individualised support and better behaviour management.

However, it was also acknowledged that resource constraints (staff time and room availability) and timetable considerations made it challenging to deliver teaching in this small group format.

Timetabling and frequency of LIT lessons

Typically schools ran the LIT programme across three to four lessons a week, with each lesson lasting 50 minutes to 1 hour. The LIT programme is designed to replace the English lessons in Year 7; while some schools adopted this approach, others chose to deliver the programme in addition to mainstream English lessons. Staff views on the strengths and weaknesses of these differing approaches are discussed here:

Replace English lessons

Schools that chose to replace English lessons with the LIT programme did so because they felt the curriculum covered by the programme was similar to that covered by mainstream English lessons and the programme could therefore replace these lessons. The advantages of this approach were felt to be that:

- It facilitated the use of dedicated English teachers to staff the programme;
- It avoided the need to remove pupils from other lessons, ensuring they had full access to the curriculum and avoiding potential resentment if pupils were withdrawn from parts of the curriculum they enjoyed:

“We felt, that the purpose of the programme was to improve their English, their literacy skills and that actually we wanted them to have the experience of a broader balanced curriculum and that actually if they had their normal kind of English lessons and then had this on top, that begins to diminish the breadth of the other subjects that they can take.”

(Assistant Principal)

- It was less stigmatising for pupils to have LIT lessons at the same time as their peers had English lessons, rather than be withdrawn from other lessons.

However, some limitations with this approach were also expressed:

- Some staff felt in an ideal world pupils would receive both the LIT programme and mainstream English lessons to consolidate their learning;
- Some concerns were raised that pupils who received the LIT programme in small groups in Year 7 may find returning to mainstream English lessons in Year 8 challenging and unfamiliar.

Run in addition to English lessons

Schools that chose to run the LIT programme in addition to mainstream English typically withdrew pupils from non-core subjects, for example Modern Foreign Languages, Technology, and Physical Education. Schools that adopted this approach felt that:

- It would ensure students progressed more quickly by increasing their exposure to English and literacy focused lessons;
- The LIT programme was not sufficient on its own to wholly replace mainstream English lessons;
- It was easier to resource because pupils could be withdrawn from classes when staff were available to teach the programme.

The limitations of this approach were felt to be that:

- Some pupils may feel stigmatised or resentful when they are withdrawn from other areas of the curriculum for additional literacy support, reducing their engagement with the programme;
- It may limit the number of English specialist teachers who can deliver the programme because they are teaching the mainstream English lessons running in addition.

Teachers' views on the programme

The main feature of the LIT programme is that it promotes use of the reciprocal teaching method alongside a set curriculum of work. This section covers participants' opinions on this specific approach and the detail of the programme in terms of what it covered, how appropriate this was for the pupils undertaking it and how well they engaged with the suggested content. Staff also made suggestions regarding the training and delivery of the LIT programme, which are summarised above.

Reciprocal teaching approach

Reciprocal teaching is an instructional activity in which students become the teacher in small group sessions. Students learn to guide group discussions using four strategies: summarizing, questioning, clarifying, and predicting. They take it in turns to take on these 'roles' within each lesson, leading a class dialogue about what has just been read in the texts. The guidance suggests that reciprocal teaching should be used consistently in each lesson of the LIT programme.

Strengths and weaknesses

Views on the reciprocal teaching approach can be broadly split into what teachers considered to be the strengths and weaknesses of the approach:

Strengths

1. **Team work and debate:** teachers felt that this approach promoted team work, independent learning as a class (without the leadership of the teacher) and open discussion. It was felt that the skills pupils gained from working this way were vital for their broader understanding within lessons:

"Discussion helps them to understand a little bit more and then they don't just think about the first answer that pops in, they do control their ideas a bit more and listen to each other to kind of find solutions."

Teachers mentioned that in order to achieve an Ofsted Outstanding rating within the school and to show 'rapid progress', there needed to be evidence of 'healthy debate' within the classroom. Teachers felt that reciprocal teaching methods facilitated this within LIT programme lessons.

2. **Enjoyment:** pupils enjoyed the roles they were given which in turn led to increased engagement and confidence in class. Teachers used different techniques such as props to distinguish between roles and pupils would swap roles frequently in order to retain engagement throughout the term.
3. **Confidence:** the approach focused on multi-sensory skills rather than just reading or writing skills and this led to increased confidence in pupils who struggled with core literacy skills. Teachers also felt that the pupils who had been previously under-confident in mainstream classes would be able to transfer the team work and communication skills gained from the approach to their mainstream classes, improving their confidence and ability to participate.
4. **Independence:** teachers described the approach as a 'flexible toolkit'. They felt it provided pupils with the skills and language to be able to both solve problems and read independently. They felt that meta-learning worked well with pupils and was a fresh approach to the norm, differing from the many other literacy programmes used with low ability pupils in primary schools:

"There's too much to learn in the world, you can't teach everyone everything but what I like about reciprocal teaching it's actually focused on teaching [them] skill...so you're actually teaching"

children how to learn as opposed to teaching them stuff [or] teaching them fact. You're teaching them skills."

Weaknesses

1. **Low ability pupils:** as outlined previously, teachers felt that pupils with the lowest levels of literacy needed more support with the programme as a whole. This was also the case with the reciprocal teaching approach. Teachers felt that these pupils struggled to understand and 'play' the roles they were given. As a result of this the lesson was slow moving and other pupils were left waiting. In particular the role of 'Questioner' was perceived to be difficult for low ability pupils:

"If they're actually going to show some sort of inference or some sort of knowledge, some sort of analysis of language or structure of the text, then they need more advanced skills."

Teachers also felt that pupils with lower ability struggled to retain the discussion or debate for long enough and greater teacher direction was required to ensure pupils were practising the necessary skills. Owing to the extra time it took to direct the class some teachers felt they could not use the approach consistently as instructed or they would not be able to cover all the content required.

2. **Repetitive lesson structures:** some teachers felt that the lesson structures were repetitive and that some pupils therefore became bored. As a result, teachers altered the lesson plans to include additional activities such as creating posters or running competitions. However, others felt that the repetitive nature of the approach was vital in reinforcing skills and that such consolidation of skills was often lacking in mainstream English lessons. Some teachers felt a conflict between ensuring consistency and making the lessons varied enough to retain pupil engagement.
3. **Establishing skills:** teachers felt it took a while for both teachers and pupils to get used to the approach. They felt pupils were used to being 'spoon fed' and that the skills required to fully participate in reciprocal teaching took time to establish and were easily forgotten by pupils during school holidays or between modules. Some teachers felt there would be value in a 'whole school' approach to reciprocal teaching to reinforce these skills and embed them across the curriculum.

Programme scope and appropriateness

Teachers praised the scope of the LIT programme, particularly in relation to the focus on reading and comprehension and the reinforcement of skills. They spoke positively about the transfer of these reading skills into other areas of the curriculum and reported that pupils began to read for pleasure in their own time. However, teachers would have liked more extended writing opportunities within the lesson plans and the programme overall to consolidate these skills. They also felt the programme would have benefited from homework tasks set alongside the LIT lesson plans.

In cases where pupils were having LIT classes and mainstream English lessons simultaneously, teachers felt that some found this confusing and that the volume of new vocabulary needed to be learnt by pupils was too ambitious.

There were also cases where there was disparity between the language used by the English department and the language used within the LIT programme (e.g. the LIT programme used the term 'language techniques' rather than 'persuasive techniques'). Similarly some teachers felt pupils were intimidated by the formal language used within the lessons, particularly language used to explain the learning objectives. Teachers felt that they would spend a 'good chunk' or 'almost half' of the lesson clarifying this language (including words such as 'modals' or 'connectors') into more simple terms. However, others felt that without this formal language and an understanding of the principles underlying the programme, pupils wouldn't be able to develop their literacy, and that this was a positive aspect of the LIT programme compared to other programmes targeted at low ability pupils.

The texts were largely felt to be age appropriate for pupils in Year 7 and teachers commented positively that the content was not patronising or 'babyish', as can sometimes be the case with texts aimed at pupils with lower levels of literacy.

While not all the content engaged every pupil, teachers were aware and understood the importance of students being exposed to texts which they would not normally choose themselves. However, there were some cultural references within the texts which teachers felt their pupils did not engage with. In particular, teachers outside London were concerned that London-centric references, such as the Underground and 'Oyster cards' were not understood by their pupils and this immediately disengaged them from the text. A similar issue was raised by teachers regarding the tarot cards mentioned in a text. It was noted that despite pupils being unsure of the meaning or context of certain cultural references, the reciprocal teaching approach provided pupils with the skills to explore the context of the stories and in some cases, continue to fully engage with the content.

Programme pace

Guidance suggests that the LIT lessons are designed to take place four times per week and that each unit will take an estimated six weeks to be delivered (including time spent on assessing pupils' progress (APP) assessment tasks). Teachers had differing views on whether the pace, as prescribed by the programme, was appropriate for their pupils. Some teachers felt that it was suitable and that they were comfortably completing the work assigned within each lesson.

Where teachers felt they could not keep up with the suggested pace, there were two main reasons cited: first, that there was too much to cover within the lesson plans; and second, they felt that by adhering to the suggested pace they would fail to meet the programme objectives and therefore prioritised meeting these objectives. Teachers felt the speed at which the programme could be delivered was driven by the ability levels of the pupils and therefore classes with more very low ability students struggled to keep up with the programme pace.

Assessment

How teachers responded to the assessment materials, training and analysis of the scores differed with experience levels. Some experienced teachers felt that the assessment structure was similar to the structures currently used by their English department and therefore found it easy to use. They also felt confident in using their own analysis to determine which literacy skills pupils needed to focus on to progress.

There was some concern that more training was needed for less experienced teachers to ensure they were able to score and then analyse and interpret these scores effectively. This concern came from both experienced teachers and less experienced teachers themselves:

"I'm still not 100 per cent sure whether or not I actually levelled it correctly...that could be something that I could have had more support with, just um, because I'm not as competent...I'm not as confident to do that."

After the interpretation of the scores, teachers held one-to-one sessions with pupils. Some teachers felt that this was difficult to organise logistically and took longer than expected although others felt it fitted in easily to the programme structure. Some teachers felt that it would be helpful to have time built into the programme to allow for identified areas of weaknesses to be addressed with pupils. It was suggested that one week of lessons was required in order to do this effectively.

Perceptions of impact

Teachers delivering the LIT programme described a range of impacts from the programme in relation to literacy, pupil engagement and enjoyment, group work skills and confidence:

Literacy

Mixed impacts were reported in relation to literacy, with examples of pupils making marked progress and other examples of limited impact. Where positive impacts were identified these were particularly related to comprehension, the ability of pupils to engage with texts independently, and the scope for the techniques taught in the programme to be used in other subject areas. The positive impacts on literacy were attributed to the following features of the programme:

- The reciprocal teaching approach
- Small group teaching
- Good resources and programme content
- Regular assessment and feedback on progress

Where pupils were felt to have stalled or were making slow progress in their literacy a number of factors were identified as barriers to impact:

- Targeting of programme and suitability to SEN, EAL and lower ability students
- Emotional and behavioural difficulties
- Staffing the LIT programme with sufficient (quality) resource

Group work skills

The small group format of the LIT programme and in particular the reciprocal teaching approach were felt to foster pupils' ability to work in groups and cooperate with their peers.

The programme was felt to promote discussion, encourage active participation and increase the ability of pupils to work independently of the teacher, all of which were viewed positively by staff delivering the programme.

Confidence

Teachers spoke positively about pupils increasing in confidence in terms of both their reading and their class participation. Again, this was felt to be facilitated by the small group sizes where pupils felt more comfortable contributing, and the reciprocal teaching approach which encouraged pupils to actively engage and participate in lessons.

Wider school impact

Teachers generally reported enjoying teaching the LIT programme, variously highlighting the small group teaching, the reciprocal teaching approach and the content of the programme as features of the programme they enjoyed. Where teachers did not enjoy teaching the programme the reason given was that its structure was too prescriptive and left limited room for adaptation and teacher input.

When asked about the feasibility of delivering the LIT programme long-term, staff fed back that the programme content complemented and worked well with the English curriculum studied in Year 7 and was consequently easy to integrate from a curriculum perspective. A number of the pilot schools were planning to continue using the programme beyond the pilot year, and some viewed it as a valuable

additional tool to support literacy in addition to other provision. The primary barrier to continued use of the programme was the cost of staffing the small group teaching as this was felt to be high and some schools struggled to resource the programme adequately.

Some teachers reported using some aspects of the LIT programme with non-treatment pupils within Year 7. In addition one school reported greater progress being made in English by pupils not receiving the programme because the pace of lessons could be speeded up without the pupils with the greatest literacy needs present. This may provide suggestive evidence for positive spillovers from the LIT programme; but this was not reported by all schools.

Conclusion

This evaluation has sought to provide new evidence on the effectiveness of an intervention that uses reciprocal teaching methods to improve the literacy skills of children in Year 7. Our analysis suggests that the LIT programme had a small positive effect on reading skills among all Year 7 pupils at the end of the intervention year – but, using our preferred analytical approach and specification, we are unable to conclude with confidence that the effect of the programme was different from zero.

It is important to note that this is *the average effect across all pupils in Year 7*, i.e. across the 15% of pupils who received the LIT programme as well as the 85% of pupils who did not. It is possible that the effect on those treated was much higher; however, it was not possible for us to produce what we consider to be an unbiased estimate of the effect of the LIT programme on this group, because the characteristics of pupils who were eligible to receive the LIT programme in treatment and control schools were simply too different for us to be confident that comparing their outcomes would yield an unbiased estimate of the impact of the programme. The estimates we did produce for this group – which focused on pupils with similar Key Stage 2 scores – were slightly bigger, but were not significantly different from zero and are very likely to be biased, as significant differences remained between pupils in treatment and control schools on a range of other characteristics that might plausibly matter for literacy progress.

We encountered several major issues in conducting our analysis. This meant that we had to rely on non-standard non-experimental methods in order to produce these impact estimates. It also meant that we were unfortunately not able to produce reliable estimates based on the primary analysis specified in the evaluation protocol (on pupils scoring in the bottom 25% on the baseline ART).

The process evaluation also suggested that there was a high degree of variability in the way in which the programme was delivered; and teachers were split about what effects they thought the programme had, what was good about it, and what didn't work so well. Unfortunately it has not been possible to identify which elements of the programme may have contributed to the small positive effect we found.

Interpretation

Our preferred specification controls for scores on the Access Reading Test (ART) taken at baseline, as well as Key Stage 1 and 2 scores. It is therefore an indication of the differential *progress* made over the course of an academic year by Year 7 pupils in schools that did and did not operate the LIT programme. These results suggest that the literacy skills (as measured by ART scores) of Year 7 pupils in schools operating the LIT programme rose 0.09 standard deviations more over the course of the year than those of Year 7 pupils in control schools. This is equivalent to one month's additional progress in reading.

However, tests indicate that this estimate is not 'statistically significant', i.e. we cannot be sure that it is different from zero. It is also worth noting that this effect is the average effect across all Year 7 pupils in treatment schools, i.e. it represents the effect among the small proportion (approximately 15%) of pupils in treatment schools who received the LIT programme and the larger proportion of pupils in treatment schools who did not. It is possible that the effect of the LIT programme on those pupils who received it was substantial; but unfortunately it was not possible for us to estimate this effect robustly, as the characteristics of pupils in treatment and control schools who were eligible to receive the LIT programme were simply too different to produce an unbiased estimate for this group.

The fact that our preferred specification produces a small positive effect – but one that is not significantly different from zero, and which we had immense trouble in estimating robustly – means that we would caution against placing too much emphasis on our results as evidence in favour of rolling out the LIT programme.

Limitations

As described above, we encountered a substantial number of issues throughout the course of the evaluation. These must be borne in mind when interpreting our results.

Testing: while the decision to test all pupils in Year 7 enabled us to estimate the impact of the LIT programme across all pupils who received it at some point during the year (plus any spillover effects onto other pupils), it also presented severe logistical problems, especially at baseline. Schools struggled to timetable all pupils to take the tests close to the start of term, meaning that pupils in the same school often took the tests several days or weeks apart. In some cases, schools also started delivering the LIT programme before they had undertaken their baseline testing.

Moreover, we found evidence of very large changes – negative as well as positive – in ART scores for some pupils between baseline and follow-up. This occurred in control schools as well as in treatment schools. This may suggest that some pupils or schools did not put maximum effort into the tests at either the start or end of the year. To the extent that this differs between treatment and control schools, this may potentially bias our estimates of the effect of the intervention.

Attrition: the intervention and evaluation teams worked closely with schools to encourage them to remain in the intervention and to conduct and return their test data on time. A substantial amount of time and effort was expended in doing so, with a relatively small number of schools dropping out of the evaluation and the vast majority of schools providing test data at both baseline and follow-up. However, schools that did drop out did not do so randomly, and those pupils which did not sit either the baseline or follow-up test or both were not a random selection of pupils in those schools either. This means that the impact estimates apply only to the subset of pupils and schools that remained in the evaluation and provided usable data.

Generalisability: the average characteristics of schools and pupils that joined and remained in the evaluation of the LIT programme, and provided usable test data, are quite different from the average characteristics of schools and pupils across England as a whole. For example, on average, 34% of pupils in schools participating in the LIT evaluation were eligible for free school meals, compared to only 16% nationally. Only 38% of pupils in our sample were of White British ethnicity, compared to 77% nationally.¹⁷ The extent to which these results could be applied to schools outside London with smaller ethnic minority populations is thus uncertain.

Estimation: we encountered severe challenges in estimating the impact of the LIT programme on children's literacy skills. This stemmed from the fact that the characteristics of pupils in treatment and control schools for whom we observed baseline and follow-up test data were unbalanced at the point of estimation; and, moreover, that they could not be satisfactorily balanced using conventional methods. This arose despite the fact that average school and pupil-level characteristics were balanced at randomization, and seems to have occurred for two main reasons:

- Drop-out of the intervention at school or pupil-level occurred non-randomly.
- As well as average differences in characteristics between pupils in treatment and control schools, we also found evidence of average differences in *combinations* of characteristics, including combinations of pupil and school characteristics. This made it especially difficult to balance the characteristics of pupils in treatment and control schools using conventional methods.

¹⁷ Table A2, <https://www.gov.uk/government/publications/gcse-and-equivalent-attainment-by-pupil-characteristics-2012-to-2013>

Instead, we had to turn to less conventional methods: looking at smaller groups of (hopefully) more similar pupils and then aggregating up to get an overall estimate of the impact of the programme. While this method enabled us to produce what we regard as a robust estimate of the impact of the programme, the statistics on which such judgements are based suggest that the balance of characteristics in our treatment and control groups was on the margins of acceptability. Thus, while we are confident that we have estimated the impact of the programme to the best of our ability, given the unbalanced nature of the data at our disposal, we would not recommend relying too heavily on this evidence alone in deciding whether or not the LIT programme should be taken up by more schools.

Feedback on the LIT programme

The feedback on the LIT programme was largely positive, with value placed on the reciprocal teaching approach, the small group nature of the teaching and the engaging and age appropriate nature of the programme content. However, a number of suggestions were made regarding ways in which the programme might be improved, which the intervention team and other schools implementing the LIT programme in future may wish to take on board:

- **Staff training and support:** the support provided by the Programme Manager was found to be very helpful, but it would clearly not be feasible to maintain the same approach at scale. An alternative suggestion was that additional support could be provided by fostering partnerships between schools delivering the programme in order to share and learn from each other's experiences. It was also suggested that a top-up training session part way through the programme would be valuable to enable staff to raise queries arising from their own teaching experience.
- **Target group:** it was generally felt that children with literacy skills just below average benefitted the most from the programme, as they had the skills necessary to engage with the content. Children with additional learning needs (such as those with special educational needs or English as an additional language) were felt to have had more difficulty accessing the programme.
- **Delivery methods:** a group of around 5–6 children seemed to be the optimum group size to which to deliver the programme. It was also felt that qualified teachers or teaching assistants with literacy experience would be best equipped to deliver the programme's content. However, there was relatively little other clear guidance regarding the best ways in which to deliver the intervention (e.g. whether it should be in addition to or instead of English lessons).
- **Programme content:** teachers fed back that they would value homework tasks being included in the programme, as well as additional content focused on basic writing skills including spelling, punctuation and sentence construction.

Lessons for future evaluations

Design:

- Ideally, evaluators of interventions at school or higher cluster levels should take into account not only the characteristics of the schools involved, but also the characteristics of the pupils who are likely to be involved in the intervention, as our results suggest that these can sometimes be very different.

Testing:

- Evaluators should be aware that the age-adjusted scores produced automatically by the ART software may censor results unnecessarily for children scoring towards the top and bottom of the distribution. The fact that 8% of children in the schools included in the LIT evaluation – which have

lower average achievement than schools throughout England – achieved the maximum score on the test suggests that it may not be very good at capturing the full dimension of ability.

- Our data suggests that some pupils' literacy skills deteriorated significantly over the course of a year, which seems unlikely and may instead indicate that some schools and/or pupils did not expend maximum effort in completing these tests. Given the reliance of many EEF evaluations on the collection of bespoke data, it might be worth considering whether schools and/or pupils could be incentivised to perform at their best on the day of the test. This is clearly controversial territory, and careful consideration would need to be given to the pros and cons of such an approach, but it is undoubtedly vital for evaluations to have access to the best possible data when estimating the effects of particular interventions.
- The merits of collecting baseline data for use in future evaluations should continue to be considered carefully. Our results suggest that the inclusion or exclusion of baseline test scores can give rise to different conclusions about the effectiveness of an intervention, over and above accounting for Key Stage 2 scores. We fully acknowledge that our evaluation is not the ideal environment from which to draw such conclusions, however, and would recommend that consideration be given to funding an experiment whose main aim (or one of them) is to understand the importance of collecting baseline data. These results could be used to provide more robust recommendations for future EEF evaluations.

Impact estimation:

- A simple comparison of means between treatment and control groups is highly unlikely to yield an unbiased estimate of the effect of the intervention, unless there has been zero (or only non-random) dropout from a randomized control trial. Pre-specified evaluation plans should always include the possibility of accounting for pupil and school characteristics if necessary. Moreover, using methods (such as propensity score matching) that produce statistics which can be used as a basis from which to judge how well the treatment and control groups are balanced can be invaluable in understanding the robustness of the results.

Future research and publications

Our results have provided some evidence that the LIT programme might be helping to raise the attainment of children who enter secondary school with below average literacy skills. However, the various challenges and uncertainties surrounding these estimates mean that we cannot draw this conclusion with any degree of certainty.

Future research could usefully continue to follow the relevant cohorts of pupils in treatment and control schools using data from the National Pupil Database in order to conduct alternative impact estimates, e.g. using Key Stage 3 or 4 scores. This would be helpful both in terms of understanding whether there is any continuing impact of the LIT programme on pupil attainment, and in providing some indicative evidence of the extent to which our results may have been biased by the use of less than ideal measures of literacy skills at the follow-up test stage.

It would also be interesting to make a direct comparison between this programme and small group literacy delivery, student vs. teacher-led learning and the role of tailored pupil assessment and feedback. All of these are features of the LIT programme, and it would be interesting to understand which of these components is the most important, or whether it is the combination that matters. It is also important to understand which delivery methods are most effective, e.g. whether it matters that it is delivered by a qualified teacher or someone who specialises in English.

We plan to write up an academic paper discussing the methods we have used to estimate the impact of the LIT programme in future. We hope to publish it in a high quality peer-reviewed journal.

References

Blackwell, M., Honaker, J. and G. King (2011) 'Multiple overimputation: A unified approach to measurement error and missing data'. Available at: <http://j.mp/jqdj72>.

Leuven, E. and B. Sianesi (2012; revised 12 Feb 2014) 'PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing', *Statistical Software Components*.

Rosenshine, B. and C. Meister (1994) 'Reciprocal teaching: A review of the research', *Review of Educational Research*, 64(4), 479–530.

Rubin, D.B (1987) *Multiple Imputation for Nonresponse in Surveys* (Wiley Series in Probability and Statistics).

Rubin, D.B (2007) 'The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials', *Statistics in Medicine*, 26(1), 20–36.

Wilson, D., Burgess, S. and A. Briggs (2011) 'The dynamics of school attainment of England's ethnic minorities', *Journal of Population Economics*, 24(2), 681–700.

Appendix A: Additional tables and figures

Table A1 School characteristics at randomization

	Control schools	Treatment schools	Difference
Expected number of pupils to get LIT	27.1	27.1	0
Total pupils on roll	1095.8	1067.5	-28.3
KS2 average point score	26.8	25.5	-1.3
% with 5 A*-C inc. English and Maths	56.6%	54.1%	-2.5%
London	73.7%	77.3%	3.6%
South East	15.8%	18.2%	2.4%
South West	10.5%	4.5%	-6.0%
Academy Converters	15.8%	18.2%	2.4%
Academy Sponsor Led	15.8%	18.2%	2.4%
Community School	36.8%	40.9%	4.1%
Foundation School	15.8%	9.1%	-6.7%
Voluntary Aided School	15.8%	9.1%	-6.7%
Voluntary Controlled School	0.0%	4.5%	4.5%
Boys school	5.3%	4.5%	-0.7%
Girls school	10.5%	18.2%	7.7%
Mixed school	84.2%	77.3%	-6.9%
% eligible for FSM	31.6%	30.0%	-1.6%
% of pupils white British ethnic origin	40.8%	39.7%	-1.2%
% of pupils Indian ethnic origin	2.9%	3.0%	0.2%
% of pupils Pakistani ethnic origin	4.7%	5.3%	0.6%
% of pupils Bangladeshi ethnic origin	9.5%	3.8%	-5.8%
% of pupils Caribbean ethnic origin	5.6%	5.7%	0.1%
% of pupils African ethnic origin	11.3%	11.6%	0.3%
% of pupils Chinese ethnic origin	0.8%	0.7%	-0.1%
% with EAL	36.0%	38.2%	2.2%
Missing % Indian ethnic origin	10.5%	13.6%	3.1%
Missing % Pakistani ethnic origin	10.5%	13.6%	3.1%
Missing % Bangladeshi ethnic origin	5.3%	0.0%	-5.3%
Missing % Caribbean ethnic origin	10.5%	9.1%	-1.4%
Missing % African ethnic origin	5.3%	0.0%	-5.3%
Missing % Chinese ethnic origin	21.1%	22.7%	1.7%
Missing % with 5 A*-C inc. English and Maths	0.0%	4.5%	4.5%

No differences are significant at the 5% level.

NatCen Social Research
is the trading name of the
National Centre for
Social Research.

Registered Office
35 Northampton Square
London EC1V 0AX
www.natcen.ac.uk

T. 020 7250 1866
F. 020 7250 1524
E. info@natcen.ac.uk
Follow us: @NatCen

A Company Limited by Guarantee
Registered in England No.4392418.
A Charity registered in England and
Wales (1091768) and Scotland (SC038454)

Table A2 School characteristics as analysed

	Control schools	Treatment schools	Difference
Expected number of pupils to get LIT	25.3	27.8	2.5
Total pupils on roll	1124.0	1078.2	-45.8
KS2 average point score	27.0	25.3	-1.8
% with 5 A*-C inc. English and Maths	56.8%	53.7%	-3.1%
London	73.3%	78.9%	5.6%
South East	13.3%	15.8%	2.5%
South West	13.3%	5.3%	-8.1%
Academy Converters	20.0%	21.1%	1.1%
Academy Sponsor Led	6.7%	15.8%	9.1%
Community School	33.3%	42.1%	8.8%
Foundation School	20.0%	10.5%	-9.5%
Voluntary Aided School	20.0%	5.3%	-14.7%
Voluntary Controlled School	0.0%	5.3%	5.3%
Boys school	6.7%	5.3%	-1.4%
Girls school	6.7%	21.1%	14.4%
Mixed school	86.7%	73.7%	-13.0%
% eligible for FSM	31.0%	29.7%	-1.3%
% of pupils white British ethnic origin	44.0%	39.1%	-4.8%
% of pupils Indian ethnic origin	2.9%	3.4%	0.5%
% of pupils Pakistani ethnic origin	4.1%	6.1%	2.0%
% of pupils Bangladeshi ethnic origin	10.3%	3.9%	-6.4%
% of pupils Caribbean ethnic origin	5.2%	5.7%	0.4%
% of pupils African ethnic origin	10.2%	11.0%	0.8%
% of pupils Chinese ethnic origin	0.8%	0.7%	-0.1%
% with EAL	35.0%	38.4%	3.4%
Missing % Indian ethnic origin	13.3%	15.8%	2.5%
Missing % Pakistani ethnic origin	13.3%	15.8%	2.5%
Missing % Bangladeshi ethnic origin	6.7%	0.0%	-6.7%
Missing % Caribbean ethnic origin	13.3%	10.5%	-2.8%
Missing % African ethnic origin	6.7%	0.0%	-6.7%
Missing % Chinese ethnic origin	20.0%	21.1%	1.1%
Missing % with 5 A*-C inc. English and Maths	0.0%	5.3%	5.3%

No differences are significant at the 5% level.

NatCen Social Research
is the trading name of the
National Centre for
Social Research.

Registered Office
35 Northampton Square
London EC1V 0AX
www.natcen.ac.uk

T. 020 7250 1866
F. 020 7250 1524
E. info@natcen.ac.uk
Follow us: @NatCen

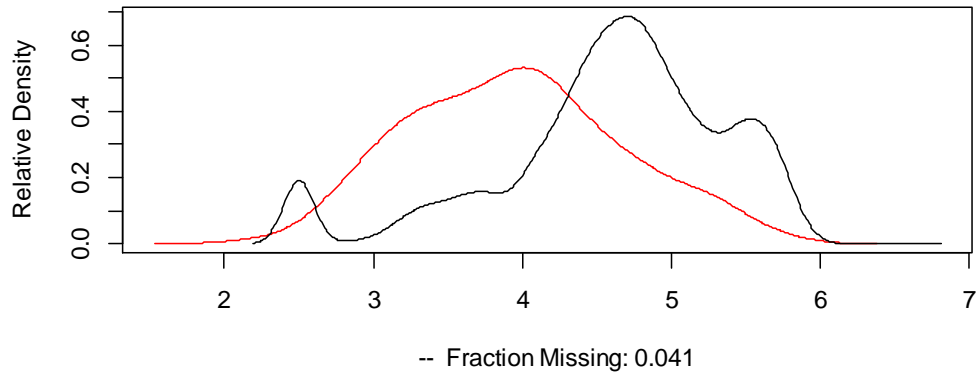
A Company Limited by Guarantee
Registered in England No.4392418.
A Charity registered in England and
Wales (1091768) and Scotland (SC038454)

Table A3 Patterns of Missingness

KS2	KS1	Female	EAL	FSM	SEN	% Bang	% Afr	Missing variables	Frequency
+	+	+	+	+	+	+	+	0	3850
+	.	+	+	+	+	+	+	1	345
+	+	+	+	+	+	.	+	1	82
.	+	+	6	70
+	+	+	+	+	+	+	.	1	19
+	+	+	5	13
.	+	+	+	5	11
.	+	+	+	+	+	+	+	1	9
.	.	+	+	+	+	+	+	2	5
+	+	+	+	4	4
+	.	+	+	+	+	.	+	2	3
.	+	+	+	+	+	.	+	2	1
.	+	.	7	1

Missing values are denoted by '.'; values that are present are denoted by '+'.

Figure A1 Observed and imputed values of KS2 scores



This figure shows the distribution of Key Stage 2 Reading scores both for pupils for whom that score was available (black) and for those for whom it had to be imputed (red). The difference between the two distributions suggests that this missingness was not at all random, and thus very important to account for when estimating the intervention's impact.

Figure A2 Distributions of age-adjusted standardised scores by test and intervention status

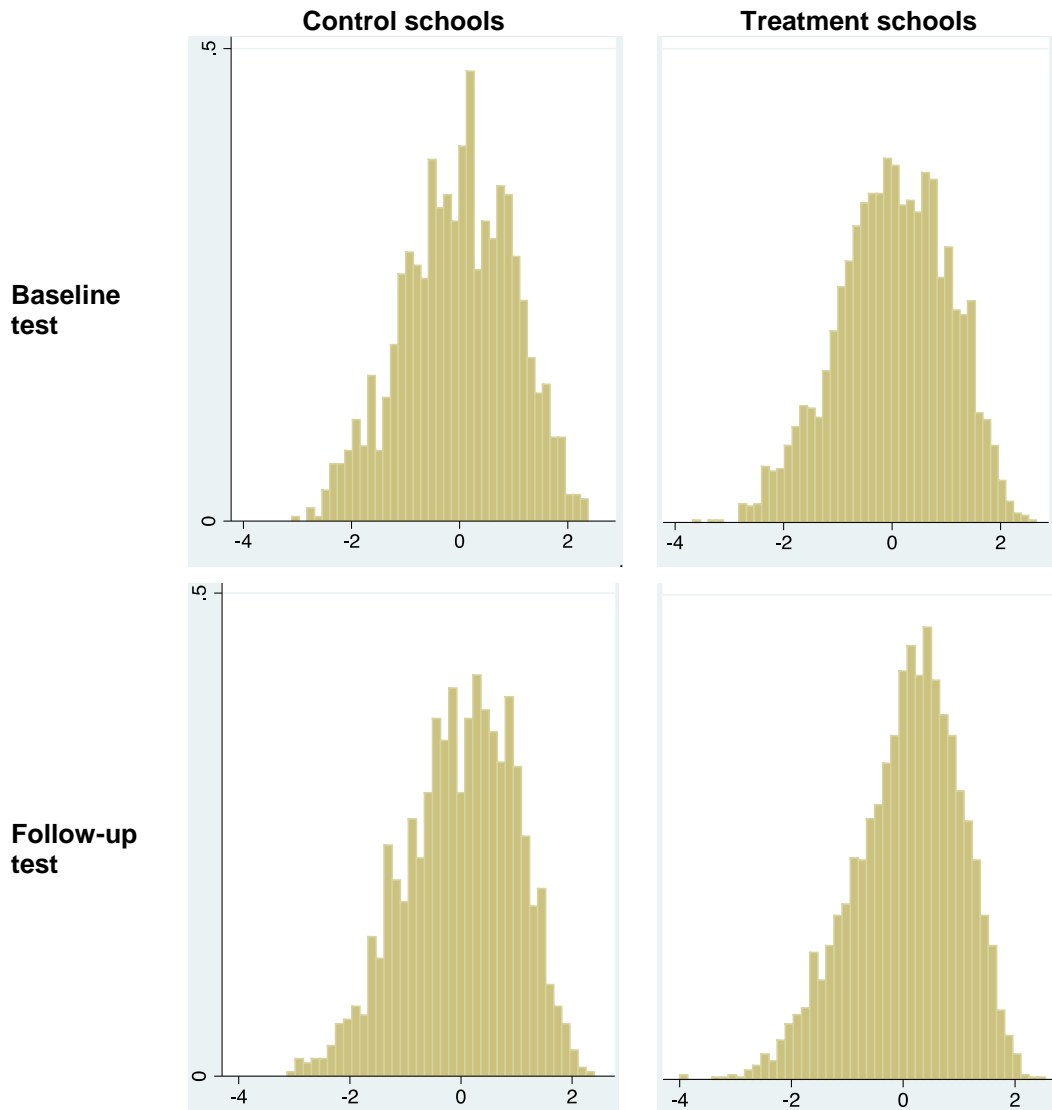
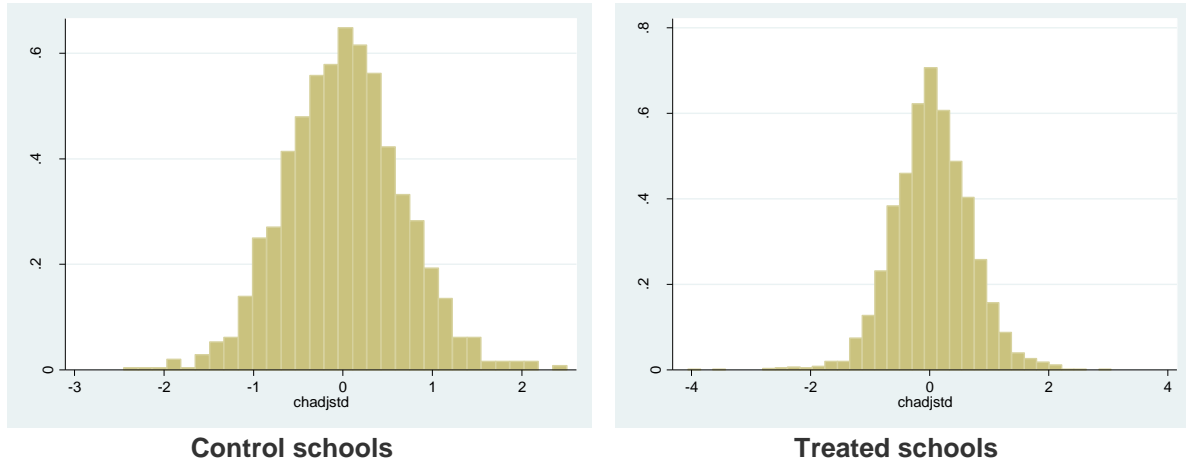


Table A4 Distribution of changes in ART scores between baseline and follow-up



This table shows the distribution of changes from the baseline to the follow-up ART test among pupils in control and treated schools, measured in standard deviations. They show that a large number of pupils saw falls in scores (i.e. negative changes), some of which were exceptionally large (>1 SD). This raises questions about whether some pupils were putting in effort in both tests.

Appendix B: Technical appendix

This appendix describes the procedures and associated diagnostics used to obtain the estimates we present of the average effect of the programme on pupils in treated schools.

Multiple imputation

As described in the main text, we restricted the analysis to those pupils for whom we observed both baseline and follow-up ART scores, and used multiple imputation to account for the uncertainty in other control variables (such as Key Stage 1 and 2 scores: see Table A3 for full details).

Multiple imputation uses variables that *are* observed to estimate a distribution (called the posterior distribution) of ‘likely’ values for each missing data point. The variance of that distribution reflects the uncertainty over the likely value.

Using these distributions, we generate five separate datasets by replacing each missing data point in the original dataset by a separate random draw from that point’s posterior distribution in each of the five imputed datasets. (We also ‘faked’ missing data – i.e. deleted values we actually observed – to test whether the multiple imputation procedure recovered them satisfactorily. It did.)

We implemented this procedure using the Amelia package in R v. 3.0.2, which is described in detail in Blackwell, Honaker, and King (2011). To obtain our final estimates, we imported into Stata the five datasets that were generated in R by Amelia. We then repeated our treatment effect analyses on each of these five datasets in turn, and combined them into a final estimate, fully incorporating all uncertainty, using Stata’s *mi estimate* command. We describe this final step in more detail below.¹⁸

Diagnostic measures

To assess precisely the balance between treated and control school pupils, taking into account combinations of characteristics, we used two diagnostic measures. The first, called *B*, is the number of standard deviations between the means of the linear propensity scores in the treated and control groups:

¹⁸ Note that all tables of school and pupil characteristics reported in the main text include original (non-imputed) values only.

$$\frac{\frac{\sum(\hat{\beta}X_{i,t})}{n_t} - \frac{\sum(\hat{\beta}X_{i,c})}{n_c}}{\left(\frac{S_t^2 + S_c^2}{2}\right)^{1/2}}$$

where $\hat{\beta}X_{i,t}$ is the linear propensity score for treated unit i and $\hat{\beta}X_{i,c}$ the same for control unit i , n_t the number of treated units, n_c the number of control units, S_t^2 the standard deviation of the propensity scores among treated units and S_c^2 the same among control units.

The linear propensity score was predicted using all pre-treatment variables (both individual- and school-level) as covariates. In an ideal randomized trial, this number should be 0: the scores should be exactly balanced.

The second diagnostic measure, called R , is the ratio of the variances of the propensity scores between treated and control groups:

$$\frac{S_t^2}{S_c^2}$$

In an ideal trial, the variances of the two groups should be the same, so their ratio should be 1.

Stratification on the propensity score

As described in the main text of the report, our data were so unbalanced that we could not rely on OLS or conventional matching procedures. Instead, we used stratification on the propensity score, as described in Rubin (2007). This works as follows:

1. Match the treated and control groups using the matching procedure that offers the best balancing for your data.

We used the Stata command PSMATCH2¹⁹ with the following configurations: default [single nearest neighbour without caliper], caliper(0.01), kernel, kernel bw(0.01), kernel kerneltype(normal), kernel bw(0.01) kerneltype(normal), radius caliper(0.01), radius caliper(0.05).

¹⁹ Leuven and Sianesi (2003).

Depending on the matching method used, and the extent to which the sample has common support – that is, the extent to which every individual in the treatment group can be matched to at least one similar-looking individual in the control group – this procedure may reduce the size of the analysis sample.

2. Re-estimate the propensity scores for the new (potentially smaller) sample.

This is necessary because if the treatment group has changed then the predicted probability of being treated (which is what the propensity score calculates) may also have changed.

3. Stratify the sample based on these new propensity scores.

If stratified into two groups (called 'bins'), they would be divided at the median re-estimated propensity score of the whole matched sample; if into three bins, then at the second and third terciles; and so on for more than three bins. In other words, all bins should have the same total (weighted) number of observations, though the proportions of treated and control units will differ between bins.

This is because the propensity score indicates how likely an individual is to be treated. In a sample such as ours where the characteristics of the treatment and control groups are very unbalanced, it can be very easy to tell which individuals fall into the treatment group, meaning that their propensity scores may all be very high, while those of the control group may all be very low. This means that there will typically be a much higher number of treatment observations in bins with high propensity scores and a correspondingly lower number in bins with low propensity scores.

Because of the size of our sample, we restricted the approach to select a maximum of 10 bins, and many fewer when we attempted to carry out subgroup analysis. (You need treatment and control units to be present in every single bin, and this does not occur with large numbers of bins in small unbalanced samples.)

4. Estimate the treatment effect separately within each bin using OLS regression.

In principle you could use any estimation procedure at this stage (more matching, a simple difference of means, etc.). Given Rubin's recommendation that B scores of 0.3 can be regarded as sufficiently well-balanced to use linear regression analysis – and, as we shall see below, that some of our bins had B values very close to this level – we chose to use OLS for our final estimation.

5. Take the average of the bin-specific effect sizes, weighting by the number of treated observations in each bin.

This produces an overall treatment effect for the programme. The weighting procedure means that treatment effects that are estimated using larger numbers of treatment individuals are given more weight than treatment effects that are estimated using smaller numbers of treatment individuals. The number of control individuals is not taken into account explicitly in this step, but – as described below – we preferred specifications which had at least a minimum number of control observations in each bin. This helps to ensure that the treatment effect is estimated robustly.

Clearly, this procedure is flexible: the number of bins can be varied, as can the matching procedure used for the first stage. The optimal combination of bins and matching procedure is simply that which jointly minimises the final weighted average values of B and R (weighting by the number of treated units in each bin, as for the final estimate of the treatment effect). Note that this is done without reference to any outcome data, let alone an estimate of the treatment effect, and that it is not only appropriate but *necessary* in order to obtain an unbiased estimate.

We considered a number of different rankings, including worst B/R across individual bins, the balance of a number of individual covariates (such as the pre-test scores), and weighted combinations of these. We concluded that our top specifications should be those with the lowest B, subject to having no bin with $B \geq 0.3$ (the maximum that Rubin suggests is acceptable for OLS estimation). We also restricted these specifications to have at least 50 control observations per bin.

The top-ranked specifications after comparing all combinations of bins (from 2 to 10) and matching procedures (mentioned above) are displayed in Table A5. The first column is the B score, and the second the R score. Next is the worst absolute value of the B score in any bin in any imputation, and likewise for the R score (presented as $|1-R|$). Next is the number of (weighted) control units in the bin with the fewest control units.

Table A5 Balancing of top specifications

	Bins	Matching Option	B	R	Worst B	Worst R	Smallest Control Bin
1	3	radius caliper(0.01)	0.09	0.52	0.28	0.65	396.47
2	3	kernel bw(0.01) kerneltype(normal)	0.10	0.77	0.24	0.66	329.47
3	3	kernel bw(0.01)	0.10	0.52	0.28	0.66	392.94
4	3	single nearest neighbour	0.19	0.82	0.43	0.42	354.00
5	3	radius caliper(0.05)	0.15	0.70	0.36	0.87	339.91
6	3	kernel	0.13	0.66	0.30	0.91	331.07
7	4	kernel kerneltype(normal)	0.31	0.75	0.52	0.59	187.66
8	3	kernel kerneltype(normal)	0.49	0.84	0.60	0.40	440.04
9	2	radius caliper(0.01)	0.54	0.70	0.72	0.48	1099.01
10	2	kernel bw(0.01)	0.54	0.71	0.71	0.50	1103.22

These specifications do a good job of balancing pupils in treated and control schools. Whereas the *B* score in the full sample was greater than 1 and the *R* score around 0.3, the final weighted *B* score in the top-ranked sample is 0.09 and the *R* score 0.52.

Full details of one specification one on one of our imputed datasets are shown in Table A6 for illustrative purposes. For each stage (i.e. each row), the columns display the *B* score, the *R* score, the treatment effect (estimated using OLS on the relevant sample in each case), the number of treated units, and the number of control units (this last number is weighted, except in row 2, which displays a count of unique units). The first row displays this information for the whole sample, the second for the matched sample (i.e. after the first round of matching has identified any treatment or control observations to drop and the propensity score re-estimated and the estimation re-run on this smaller sample).

Table A6 Illustrative matching process (top specification, imputation 2)

	B	R	Treatment effect	Number of treated units	Number of control units
Whole Sample	1.06	0.30	0.12	2889	1524
Matched Sample	0.56	1.68	0.11	2804	1150
Bin-1	0.28	0.35	0.25	908	961.34
Bin-2	0.12	0.97	0.07	424	1446.19
Bin-3	-0.03	0.45	-0.01	1472	396.47
Weighted Avg.	0.12	0.50	0.09	2804	2804

Individual bins follow. The last row displays the weighted averages of the individual bins, with the exception of the number of treated and control units, for which totals are displayed. The figures in this final row suggest it is now appropriate to rely on regression adjustment to obtain the treatment effect. Indeed, it can now be seen that OLS regression on the whole (unmatched) sample using this specification in this imputation would yield an estimate that is slightly too high.

To obtain our final estimate, we took the weighted average treatment effect estimated using each of the five imputed datasets individually and averaged them (i.e. we averaged the figures in the final row across each of our five imputed datasets).

We used a bootstrap procedure to estimate standard errors.²⁰ For each imputation, we generated 100 bootstrap samples using Stata's *bootstrap* command, blocking at the school level. For each of these samples, we estimated a treatment effect using the full stratification process described above. These treatment effects together yielded the standard error for that imputation. To produce the final overall result, we combined the bootstrapped standard errors for the different imputations using the formula

$$T = \sqrt{\frac{1}{m} \sum_{j=1}^m U_j + \left(1 + \frac{1}{m}\right) \left(\frac{1}{m-1}\right) \sum_{j=1}^m (\hat{Q}_j - \bar{Q})^2}$$

where the red expression is the within-imputation variance and the blue expression the between-imputation variance.²¹ In particular, \hat{Q}_j is the regression coefficient obtained from dataset j of m total imputed datasets; U_j is the standard error associated with \hat{Q}_j , and \bar{Q} is the average of the \hat{Q}_j s).

Table A7 displays the top-ranked specifications with effect sizes and standard errors. (These three are in fact the only specifications for which B never rises above 0.3 in any bin.) It can be seen that the result of a treatment of around 9% of a standard deviation which is not significant at the 5% level is robust to the choice between our top specifications. This gives us additional confidence in our results. Moreover, this result is very different (in terms of significance level) to that obtained using OLS on the raw sample (although the point estimates are reasonably close).

Table A7 Results from top specifications

Specification	Estimate (Standard error)	P-value	95% confidence interval
1	0.09 (0.07)	0.16	(-0.04, 0.22)
2	0.09 (0.07)	0.19	(-0.05, 0.23)
3	0.08 (0.07)	0.28	(-0.06, 0.22)
OLS	0.12 (0.04)	<0.01	(0.04, 0.20)

²⁰ It is not yet known whether bootstrapping is appropriate for this estimator; however, the analysis team conducted simulation studies to investigate its finite sample properties and concluded that this approach was both defensible and the best option available.

²¹ This formula was proposed in Rubin (1987).

Standard errors are clustered at the school level.

The impact of the LIT programme on pupils eligible to receive it

The original plan was for the primary analysis to estimate the impact of the LIT programme on pupils who were eligible to receive it, i.e. on those scoring at or below the 25th percentile on the baseline ART. When we ran specification searches restricted to such pupils, however, the resultant statistics signified that we should not rely on the estimates produced by even our best specification.

For example, while our best specification had a final averaged B score of 0.27 (just on the threshold of acceptability), it had a worst B score of 0.59 (very far from acceptable). The bin with the smallest number of control observations was also very small indeed: just four individuals. Moreover, the bins with the smallest number of control observations were also the bins with the largest number of treated observations (because the bins are grouped according to propensity score, and there is a negative correlation between the propensity scores assigned to treatment and control group observations). This means that the within-bin estimates given the heaviest weight in calculating the overall assessment of the impact of the programme on this group has been estimated using a control group containing just four individuals. For comparison, the equivalent figures for the whole sample are an averaged B score of 0.09, a single worst B score of 0.28, and a smallest control bin of 396.

These statistics highlight that any estimate of the treatment effect calculated using this sample would likely be utterly misleading. Indeed, as Table A8 shows, the impact of the LIT programme on pupils scoring below the 25th percentile on the baseline ART (the group of pupils who should have received the intervention) fluctuates wildly depending on the specification, ranging from as low as 0.05 to as high as 0.17. Moreover, across all possible specifications, the lowest ATT was -0.78 (a negative effect of minus three quarters of a standard deviation) and the highest was 0.54 (a positive effect of half a standard deviation). Given this range, it would seem unwise to read anything into these estimates.

Table A8 Results and balancing of top specifications in ITT analysis of bottom 25%

	Bins	Matching Option	ATT	B	R	Worst B	Worst R	Smallest Control Bin	Treated individuals in SCB
1	6	radius caliper(0.05)	0.17	0.27	1.00	0.59	0.79	4.14	215
2	8	radius caliper(0.01)	0.12	0.27	1.27	1.32	3.16	2.53	142
3	6	kernel kernel	0.12	0.28	1.03	0.67	0.79	4.06	215
4	8	bw(0.01) radius	0.08	0.28	1.27	1.26	9.33	2.52	142
5	9	caliper(0.01)	0.05	0.30	1.21	1.25	3.77	1.66	121

It is important to account for differences between treatment and control groups in terms of any characteristics that might matter for pupil progress. However, it is a judgement call as to which characteristics matter most. We strongly felt that it was necessary to account for differences in a range of pupil and school characteristics in order to produce unbiased estimates of the impact of the LIT programme; however, some characteristics may matter more than others. For example, it might

be reasonable to suppose that pupils' baseline test scores are the most important factor in predicting how much progress they are likely to make over the coming year.

With this in mind, we additionally produced estimates of the impact of the LIT programme on pupils who received it by comparing pupils who scored in the bottom 25% on the ART pre-test in treatment and control schools. The results and diagnostics for the top 5 specifications are shown in Table A9.

All the B scores are close to zero, as we would expect given that they represent the balancing of a single covariate for which we are controlling. These low B scores merely indicate that, if balancing on Key Stage 2 scores is all we care about, it would be appropriate to rely on linear regression adjustment. Indeed, we obtain an almost identical effect size (0.12) from the equivalent OLS specification. However, as can be seen from the rightmost column, which shows B scores calculated using all covariates (and so comparable with all other B scores in this report), the treatment and control groups here are severely unbalanced. It is therefore likely that the effect size reflects the differing composition of the two groups as much as any effect of the programme. This worry is affirmed by the differing estimates obtained when we do try to account for these other characteristics (as in Table A8). We therefore do not believe any of the effect sizes we report for this bottom 25% group to be unbiased, and caution strongly against making any use of them.

Table A9 Results and balancing of top specifications in ITT analysis of bottom 25%, controlling only for Key Stage 2 scores

	Bins	Matching Option	ATT	B	R	Worst B	Worst R	Smallest Control Bin	B with all covariates
1	2	caliper(0.01)	0.13	0.00	0.97	0.01	0.96	310	1.49
2	3	caliper(0.01)	0.13	0.00	0.96	0.01	0.93	207	1.45
3	2	single NN	0.13	0.00	0.98	0.01	0.96	310	1.50
4	3	single NN	0.13	0.00	0.96	0.01	0.93	209	1.46
5	4	caliper(0.01)	0.13	0.01	0.94	0.02	0.88	150	1.57

Subgroup analysis

We encountered similar issues to those described above when we tried to estimate the impact of the LIT programme on various subgroups of pupils, including those eligible for free school meals. Table A10 shows the top specifications for the FSM subgroup. It can be seen that the B scores are well beyond the threshold of acceptability in every case, while the estimated ATT varies wildly. Again, it would seem unwise to read anything into any of these numbers.

Table A10 Results and balancing of top specifications in subgroup analysis

	Bins	Matching Option	ATT	B	R	Worst B	Worst R	Smallest Control Bin	Treated individuals in SCB
1	4	caliper(0.01)	0.12	0.38	1.77	1.02	3.75	71	300
2	10	caliper(0.01)	-2.45	0.40	2.00	1.78	17.91	7	141
3	3	kernel	0.08	0.40	2.12	0.89	2.81	142.88	348

		bw(0.01)							
		radius							
4	3	caliper(0.01)	0.08	0.40	2.26	0.93	3.28	145.53	346
5	5	caliper(0.01)	0.47	0.41	1.78	1.18	3.63	51	245

Effect on estimates of controlling for baseline ART scores

The estimate we report as our primary result in the main text includes baseline ART scores as a control variable. Given the concerns raised over the quality of the ART test, however, we also repeated the analysis excluding baseline ART scores as a robustness check. We found that *not* controlling for these baseline scores gave rise to a larger and significant estimate of the treatment effect in all specifications we tested. Moreover, regressing baseline test scores on our treatment indicator, and controlling for all other covariates, including Key Stage scores, yielded a large (10% of a standard deviation) and significant treatment ‘effect’. This could be for one of (at least) two reasons:

1. It could be that Key Stage 2 scores do not fully capture pupils’ abilities in the dimensions measured by the ART, and treated school pupils really were of higher ability than control school pupils on these important dimensions. Thus, *not* controlling for baseline ART test scores would cause us to *overestimate* the treatment effect.
2. Alternatively, it could be that treatment schools were more enthusiastic about implementing the baseline test. (Schools knew their treatment assignment at the time of testing, and the intervention delivery co-ordinator supported them in administering the test. Control schools received no such support.) In this case, pupils of a given ability in treatment schools would have scored more highly than children of the same ability in control schools. As such, controlling for baseline ART test scores would lead us to *underestimate* the treatment effect.

Both of these explanations are plausible. Indeed, they could both be true. If they are, then the actual effect size would lie between the estimate that did control for pre-test score (9% of a standard deviation) and the estimate that did not (22% of a standard deviation). It is worth noting here that the upper bound of the confidence interval for our primary result is 23%, while the lower bound of the confidence interval for this alternative estimate is 8%, so the estimates are not ‘inconsistent’ with one another. The sizeable discrepancy does, however, raise the question of whether Key Stage test scores are the right metrics to use as baseline measures of attainment in some or all cases, or whether it is safer to collect separate baseline measures of attainment for each individual evaluation. Further investigation of this important issue would be worthwhile.

Appendix C: Power calculations from the protocol

The tables below show the total sample required (number of children across programme and control schools) for 80% statistical power. Within each table, the required sample varies according to the estimated effect size (measured in standard deviations) and the within-school correlation in test scores.²² Each table assumes a different variance for the test scores: for a simple t-test using standardised scores, the variance will be 1. The *residual* variance will be lower if a regression is fitted using other explanatory variables (or baseline measures from the pre-test). Tables 2, 3, and 4 assume that the regression model explains 25%, 50%, and 75% respectively of the variance of the test score.

Table 1. Variance of test score = 1

		Within-school correlation			
		0	0.1	0.2	0.3
Effect size	0.1	3,136	N/A ²³	N/A	N/A
	0.2	784	N/A	N/A	N/A
	0.3	348	4,213	N/A	N/A
	0.4	196	368	N/A	N/A
	0.5	125	169	301	219,520

Table 2. Residual variance of test score = 0.75

		Within-school correlation			
		0	0.1	0.2	0.3
Effect size	0.1	2,352	N/A	N/A	N/A
	0.2	588	N/A	N/A	N/A
	0.3	261	769	N/A	N/A
	0.4	147	217	537	N/A
	0.5	94	113	150	263

Table 3. Residual variance of test score = 0.5

		Within-school correlation			
		0	0.1	0.2	0.3
Effect size	0.1	1,568	N/A	N/A	N/A
	0.2	392	N/A	N/A	N/A
	0.3	174	292	1,872	N/A
	0.4	98	119	164	313
	0.5	63	68	75	88

Table 3. Residual variance of test score = 0.25

		Within-school correlation			
		0	0.1	0.2	0.3
Effect size	0.1	784	N/A	N/A	N/A
	0.2	196	368	N/A	N/A
	0.3	87	102	130	199
	0.4	49	51	53	56
	0.5	31	31	30	29

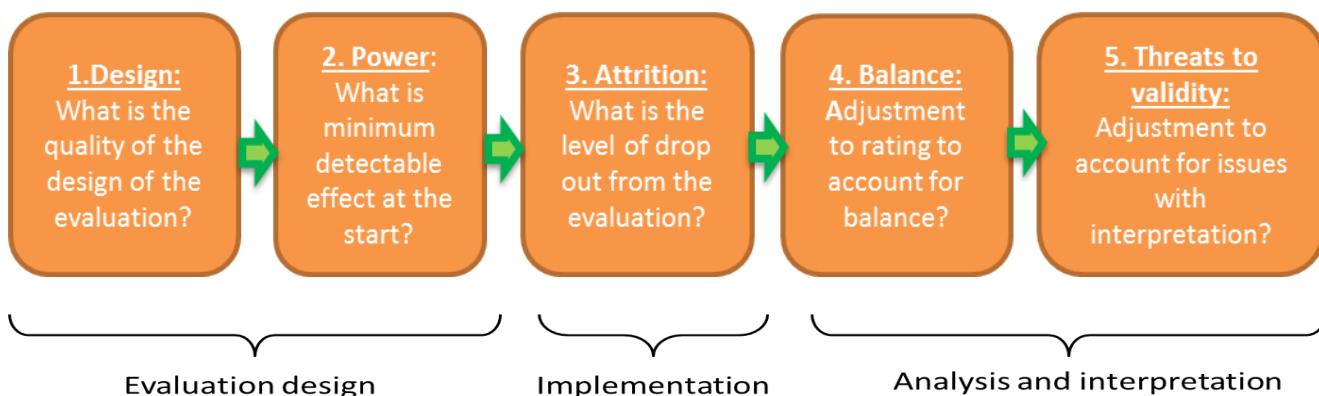
Where detection at 80% power is feasible, most of the numbers are under 1,000, meaning that there is a very good chance that an overall impact can be detected given that the evaluation is following 40 schools. Under some circumstances, it will also be possible to detect impacts for sub-

²² The power calculations also assume a significance level of 5%, 40 schools (clusters), split equally between treatment and control groups, and a coefficient of variation for the cluster size of 0.25.

²³ 'N/A' in a particular cell means that the number of clusters (schools) is insufficient for the power required, given the effect size and within-school correlation.

groups as well. Under a basic t-test, 368 pupils would be required in the sample to detect an impact of 0.4 standard deviations (with 80% power) if the within-school correlation is 0.1. This points to sub-group analysis being feasible for approximately three groups if around 1,000 pupils are selected in total. If a regression model is used which explains 25% the variance of the test score (Table 2), the required samples are smaller and approximately five sub-groups more could be analysed under the same assumptions.

Appendix D: EEF Security Rating Summary



Rating	1. Design	2. Power (MDES)	3. Attrition	4. Balance	5. Threats to validity
5	Fair and clear experimental design (RCT)	< 0.2	< 10%	Well-balanced on observables	No threats to validity
4	Fair and clear experimental design (RCT, RDD)	< 0.3	< 20%	↓	↓
3	Well-matched comparison (quasi-experiment)	< 0.4	< 30%	↓	↓
2	Matched comparison (quasi-experiment)	< 0.5	< 40%	↓	↓
1	Comparison group with poor or no matching	< 0.6	< 50%	↓	↓
0	No comparator	> 0.6	> 50%	Imbalanced on observables	Significant threats

The final security rating for this trial is 1 . This means that findings are of low security.

The trial was designed as a cluster randomized efficacy trial with the intention of recruiting 40 schools. Only 34 were achieved providing a minimum detectable effect of 0.25 at randomization for the subgroup of pupils eligible for LIT (and 0.16 for all pupils) meaning the trial could still have achieved a maximum of 4 . However, the trial experienced high attrition with 21% of schools dropping out. In addition the sample of pupils used in the analysis was significantly imbalanced at the baseline on FSM, ethnicity and prior attainment. The evaluators used matching techniques to achieve balance for all pupils but could not achieve this for the 15% of pupils who received the intervention. This means the estimate is for all pupils in Year 7 and any effect on the treated would have been diluted and the result is difficult to interpret. In addition the tests were delivered by schools under exam conditions, so there is no guarantee that staff did not interfere with test delivery.

Come and read our
latest blog

natcenblog.blogspot.com

Appendix E: Consent letters

1. Treatment Schools «C1firstname»

31st August 2012

«C1lastname»

«SchName»

Our ref: P10053 / «Serial»

«Add1»

«Add2»

«Add3» «Add4»

Dear «C1firstname» «C1lastname»,

RE: Evaluation of the LIT Reading Programme

Thank you very much for agreeing to take part in the evaluation of the LIT Reading Programme. This letter provides information about what is involved and answers key questions about the study. The Education Endowment Foundation is funding the programme along with its evaluation which is being carried out by the Institute for Fiscal Studies and NatCen Social Research.

When is my school delivering LIT?

Forty schools in total are taking part in the programme: 20 schools will deliver LIT to eligible pupils during 2012-13 and 20 schools will offer the programme in 2013-14. The second group of schools will act as a comparison this year so that we can find out how LIT impacts on Year 7 pupils' reading ability. Your school is delivering LIT in 2012-13 and therefore is an 'intervention' school this year.

What does the evaluation involve?

Data for the evaluation will be collected through the following activities:

1. Testing the reading ability of all Year 7 pupils with the Access Reading Test in September 2012 and June/July 2013.
2. Carrying out a short Single Word Reading Test with the pupils selected for LIT in September 2012 and June/July 2013.
3. A visit from NatCen researchers to a small number of schools to observe the administration of the reading test in September 2012.
4. Completing a short (5 minute) online survey about reading support in Year 7 in the Autumn term 2012.
5. Interviews with LIT staff in some of the schools to find out more about how LIT is being delivered and their views on the programme.

This letter provides details about the tests (1 and 2). We will be in touch separately about the other research activities listed above.

NatCen

Social Research that works for society

Do pupils have to take part?

No. Although we would like as many pupils as possible to participate, for ethical reasons, parents/carers need to be given the opportunity to withdraw their child from the evaluation. We are sending you 'opt out' letters to distribute to the parents/carers of all Year 7 pupils as soon as possible. The letters provide information about the evaluation and ask parents to tell the school office within a week if they don't want their child to take part in the tests. After a week, all pupils whose parents have not requested to opt out will take part in the tests.

What do I need to do and when?

	Task	When
1	Give out parent opt-out letters to all Year 7 parents/carers	This week (first week of term)
2	All Year 7 pupils who have not opted out take part in the Access Reading Test (ART) developed and provided by Hodder Education	As soon as possible after the opt-out period (i.e. in the second week of term if possible)
3	Select pupils for LIT if they score at or below the 25 th percentile on the ART	When ART is complete for all Year 7 pupils
4	Send the full ART data, along with pupil UPNs, to the Institute for Fiscal Studies	When ART is complete for all Year 7 pupils
5	Carry out Single Word Reading Test with pupils selected for LIT	After LIT pupils are selected and before LIT begins
6	Send the Single Word Reading Test to NatCen Social Research	When complete with all LIT pupils
7	Repeat the ART and Single Word Reading Test and send the test results to IFS and NatCen again	June/July 2013 (we will be in touch to arrange this)

What do I need to know about the Access Reading Test (ART)?

The ART is completed on computers and takes around 30 minutes. Please use **Version A** of the test.

Hodder Education will provide your school with the test in the form of a CD-ROM in early September. We advise that your school's network administrator installs the programme installation and oversees the set-up.

Pupil information (full name and Unique Pupil Number) can be keyed into the programme before the test is taken by pupils, or imported into the system from your school management system. The file formats the programme is able to import are .XLS, .CSV, and .TXT files. If importing pupil information from a database/management system, the file needs to have the field names (i.e. the headings of the data columns) as the *first* line in the file or spreadsheet for the data import to work.

It is really important that each pupil's UPN is included because we need to link in background information about the pupil from the National Pupil Database. Please insert UPN in the same field as pupil name: *Firstname Lastname UPN*. (Although there is a separate field for UPN, this is not imported with the results data following the completion of the test.)

Please enter the name of your school and URN (unique school number) as follows: Schoolname (URN) e.g., Woodgreen School (159386).

Produce the "results list" at individual pupil level (as opposed to class or cohort level) which can be done from within the Admin area of the programme.

- Save the file. Use Winzip to encrypt and password protect the file.
- Email the ART results to art.results@ifs.org.uk (a secure inbox which has been set up solely to receive this data). Please phone Haroon Chowdry at IFS to give the password 0207 291 4800.

What do I need to know about the Single Word Reading Test (SWRT)?

Please complete this test with the pupils selected for LIT before the programme begins.

This is a short pen & paper test that takes about five minutes to complete.

Hodder Education is going to provide your school with paper copies of the test.

Write on the pupil name and UPN before the test is taken by the relevant pupils.

Write on the school name and URN to each test paper.

NatCen will arrange the secure collection of the completed SWRT papers via a courier.

Where can I get technical support with the tests?

The *Access Reading Test Manual* provides a good and clear overview of how to install and use the test. We will send it by email at the start of term.

Network administrators may find the information at the website below useful:
[http://msdn.microsoft.com/en-us/library/cb6t8dtz\(VS.80\).aspx](http://msdn.microsoft.com/en-us/library/cb6t8dtz(VS.80).aspx)

If you still have technical questions, please contact Hodder on:
support@candlservices.co.uk or telephone at 01275 541253.

Where can I get more information about the Single Word Reading Test or the evaluation?

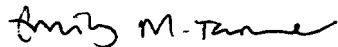
Please contact NatCen on freephone 0800 652 0401 leaving your telephone number so that we can call you back or email us on LIT@natcen.ac.uk.

Will the data be treated securely?

The results from the tests will be analysed anonymously by the Institute for Fiscal Studies and data about pupil characteristics will be added from the National Pupil Database, maintained by the Department for Education. The Institute for Fiscal Studies is registered under the Data Protection Act and any data it receives will be treated as strictly confidential, in accordance with the Act.

Thank you very much for taking part in this trial. This research will provide valuable insights into the effectiveness of the LIT Programme for the benefit of pupils both now and in the future.

Yours sincerely,



Emily Tanner
Senior Research Director, NatCen Social Research

Come and read our
latest blog

natcenblog.blogspot.com

2. Control Schools

«C1firstname» «C1lastname»

31st August 2012

«SchName»

«Add1»

Our ref: P10053 / «Serial»

«Add2»

«Add3» «Add4»

«Pcode»

Dear «C1firstname» «C1lastname»,

RE: Evaluation of the LIT Reading Programme

Thank you very much for agreeing to take part in the evaluation of the LIT Reading Programme. This letter provides information about what is involved and answers key questions about the study. The Education Endowment Foundation is funding the programme along with its evaluation which is being carried out by the Institute for Fiscal Studies and NatCen Social Research.

When is my school delivering LIT?

Forty schools in total are taking part in the programme: 20 schools will deliver LIT to eligible pupils during 2012-13 and 20 schools will offer the programme in 2013-14. The second group of schools will act as a comparison this year so that we can find out how LIT impacts on reading ability. Your school is delivering LIT in 2013-14 and therefore is a 'control' school this year.

What does the evaluation involve?

Data for the evaluation will be collected through the following activities:

1. Testing the reading ability of all Year 7 pupils with the Access Reading Test in September 2012 and June/July 2013.
2. A visit from NatCen researchers to a small number of schools to observe the administration of the reading test in September 2012.
3. Completing a short (5 minute) online survey about reading support in Year 7 in the Autumn term 2012.
4. Interviews with teaching staff in some of the schools to find out more about literacy support for Year 7 pupils.

This letter provides details about the tests (1). We will be in touch separately about the other research activities listed above.

Do pupils have to take part?

No. Although we would like as many pupils as possible to participate, for ethical reasons, parents/carers need to be given the opportunity to withdraw their child from the evaluation. We are sending you 'opt out' letters to distribute to the parents/carers of all Year 7 pupils as soon as possible. The letters provide information about the evaluation and ask parents to tell the school office within a week if they don't want their child to take part in the tests. After a week, all pupils whose parents have not requested to opt out will take part in the tests.

What do I need to do and when?

	Task	When
1	Give out parent opt-out letters to all Year 7 parents/carers	This week (first week of term)
2	All Year 7 pupils who have not opted out take part in the Access Reading Test (ART) developed and provided by Hodder Education	As soon as possible after the opt-out period (i.e. in the second week of term if possible)
3	Send the full ART data, along with pupil UPNs, to the Institute for Fiscal Studies	When ART is complete for all Year 7 pupils
4	Repeat the ART and send the test results to IFS again	June/July 2013 (we will be in touch to arrange this)

What do I need to know about the Access Reading Test (ART)?

The ART is completed on computers and takes around 30 minutes. Please use **Version A** of the test.

Hodder Education will provide your school with the test in the form of a CD-ROM in early September. We advise that your school's network administrator installs the programme installation and oversees the set-up.

Pupil information (full name and Unique Pupil Number) can be keyed into the programme before the test is taken by pupils, or imported into the system from your school management system. The file formats the programme is able to import are .XLS, .CSV, and .TXT files. If importing pupil information from a database/management system, the file needs to have the field names (i.e. the headings of the data columns) as the *first* line in the file or spreadsheet for the data import to work.

It is really important that each pupil's UPN is included because we need to link in background information about the pupil from the National Pupil Database. Please insert UPN in the same field as pupil name: *Firstname Lastname UPN*. (Although

there is a separate field for UPN, this is not imported with the results data following the completion of the test.)

Please enter the name of your school and URN (unique school number) as follows: Schoolname (URN) e.g., Woodgreen School (159386).

Produce the “results list” at individual pupil level (as opposed to class or cohort level) which can be done from within the Admin area of the programme.

- Save the file. Use Winzip to encrypt and password protect the file.
- Email the ART results to art.results@ifs.org.uk (a secure inbox which has been set up solely to receive this data). Please phone Haroon Chowdry at IFS to give the password 0207 291 4800.

Where can I get technical support with the tests?

The *Access Reading Test Manual* provides a good and clear overview of how to install and use the test. We will send it by email at the start of term.

Network administrators may find the information at the website below useful:
[http://msdn.microsoft.com/en-us/library/cb6t8dtz\(VS.80\).aspx](http://msdn.microsoft.com/en-us/library/cb6t8dtz(VS.80).aspx)

If you still have technical questions, please contact Hodder on:
support@candlservices.co.uk or telephone at 01275 541253.

Where can I get more information about the evaluation?

Please contact NatCen on freephone 0800 652 0401 leaving your telephone number so that we can call you back or email us on LIT@natcen.ac.uk.

Will the data be treated securely?

The results from the tests will be analysed anonymously by the Institute for Fiscal Studies and data about pupil characteristics will be added from the National Pupil Database, maintained by the Department for Education. The Institute for Fiscal Studies is registered under the Data Protection Act and any data it receives will be treated as strictly confidential, in accordance with the Act.

Thank you very much for taking part in this trial. This research will provide valuable insights into the effectiveness of the LIT Programme for the benefit of pupils both now and in the future.



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

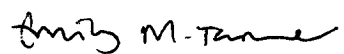
21–24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk

Yours sincerely,



Emily Tanner
Senior Research Director, NatCen Social Rese



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

21-24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk

Year 7 Parent / Carer

3rd September 2012

Dear Parent / Carer,

RE: Evaluation of the LIT Reading Programme

Your child's secondary school has chosen to take part in an exciting project to test the effectiveness of a literacy programme called 'LIT' for raising the reading ability of Year 7 pupils in need of additional support. The Education Endowment Foundation is funding the programme and its evaluation which is being carried out by the Institute for Fiscal Studies and NatCen Social Research.

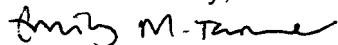
Forty schools in total are taking part in the programme: 20 schools will deliver LIT to eligible pupils during 2012-13 and 20 schools will offer the programme a year later. As part of the evaluation, all Year 7 pupils will take part in an online reading test lasting approximately 30 minutes during September 2012 and again in Summer 2013. The results will be used to select pupils for the LIT programme and to establish the impact of the programme on literacy development.

The results from the tests will be analysed anonymously by the Institute for Fiscal Studies and data about pupil characteristics will be added from the National Pupil Database, maintained by the Department for Education. The test results will be treated as strictly confidential, in accordance with the Data Protection Act.

Providing effective reading support to the pupils who need it is extremely important for ensuring that children achieve their potential. This research will provide valuable insights into the effectiveness of the LIT Programme for the benefit of pupils both now and in the future.

We do hope that you are willing for your child to take part. If you are not, please let your school administrator know within a week. If you have any questions about the study, please email us on LIT@natcen.ac.uk or call freephone 0800 652 0401 leaving your telephone number so that we can call you back.

Yours sincerely,



Emily Tanner
Senior Research Director, NatCen Social Research

Appendix F: Cost Rating

Cost ratings are based on the approximate cost per pupil of implementing the intervention over one year. Cost ratings are awarded using the following criteria.

Cost	Description
£	<i>Very low:</i> less than £80 per pupil per year.
£ £	<i>Low:</i> up to about £170 per pupil per year.
£ £ £	<i>Moderate:</i> up to about £700 per pupil per year.
£ £ £ £	<i>High:</i> up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per year.



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

21–24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk

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 v2.0.

To view this licence, visit www.nationalarchives.gov.uk/doc/open-government-licence/version/2 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 www.educationendowmentfoundation.org.uk.



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

21–24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk